
Obsah

Úvod	7
1 Automatické indexování textů	15
1.1 Význam pojmu „indexování“ v relačních databázích, v textových databázích a v jiných podobných systémech	15
1.2 Problém rozpoznání (ohodnocení) významnosti jednotlivých výrazů v textu a možnosti jeho řešení	17
1.2.1 Možnosti dichotomického výběru	17
1.2.2 Metody vážení selekčních znaků	20
1.3 Problém tvarosloví (morfologie) přirozeného jazyka a možnosti jeho řešení	25
1.3.1 Možnosti a meze použití operátoru pravostranného rozšíření	26
1.3.2 Základní pojmy morfologie	27
1.3.3 Derivátor slovních tvarů	30
1.3.4 Lematizátor slovních tvarů	43
1.3.5 Zobecnění derivátoru, resp. lematizátoru	57
1.3.6 Jiné možnosti řešení problému tvaroslovné a pravopisné variability	58
1.4 Možnosti rozpoznání koherentních víceslovných výrazů	62
1.4.1 Metody založené na (omezené) syntaktické analýze	62
1.4.2 Statistické metody rozpoznávání koherentních slovních spojení . .	64
1.5 Automatické indexování pomocí tezauru	65
1.5.1 Obecná motivace	65
1.5.2 Základní definice tezauru a možnosti jeho počítačové reprezentace	66
1.5.3 Problémy formalizace indexování podle tezauru	70
1.5.4 Tezaurus jako pomůcka při formulaci dotazů	76
1.6 Automatické indexování pomocí elementárních sémantických znaků	78
1.7 Problém homonymie v přirozených jazycích	81
1.7.1 Morfologická homonymie	81

1.7.2	Obecná homonymie slov	81
1.7.3	Homonymie selektivně významného a nevýznamného slova	83
1.7.4	Homonymie odborného termínu s neterminologickým užitím formálně stejného výrazu	84
1.7.5	Homonymie grafických prvků interpunkčního typu	85
1.7.6	Nárůst homonymie v „telegrafické“ češtině	86
1.7.7	Homonymie syntaktických struktur	87
1.7.8	Obecná „řešení“ problému homonymie	92
2	Možnosti automatizace tvorby tezaurů	95
2.1	Databázový systém jako prostředí pro tvorbu a údržbu tezauru	95
2.2	Extrakce terminologické slovní zásoby z textů	97
2.3	Automatizace vyhledávání tezaurových vztahů mezi termíny	98
2.3.1	Vyhodnocování závislostí mezi výskyty termínů v dokumentech	98
2.3.2	Porovnávání termínů na základě metody <i>SĚMAN</i> (princip systému <i>ATEZ</i>)	100
2.3.3	Rozpoznávání tezaurových vztahů na základě morfosyntaktických vzorců	103
3	Automatické referování (vytváření abstraktů, anotování, sumarizace)	105
3.1	Co je cílem této činnosti	105
3.2	Obecné zásady	107
3.3	IBM Summarization tool	107
3.4	Automatické referování založené na měření obsahových souvislostí mezi větami	108
3.5	Doporučovaná obecná vylepšení metod výběru vět s nejvyšší vahou	113
3.6	Metoda detekce indikátorů klíčových sdělení	114
4	Strojový překlad	117
4.1	Východiska a počátky strojového překladu	117
4.2	Strojový překlad 1. generace	119
4.2.1	Překlad typu „slovo za slovo“	119
4.2.2	Hlavní důvody nemožnosti korektního překladu metodou „slovo za slovo“	121
4.2.3	První pokusy o překlad typu „věta za větu“	127
4.2.4	Problém označování generací strojového překladu	128
4.3	Idea strojového překladu 2. generace	129
4.3.1	„Ideální schéma“ a diskuse k němu	129
4.3.2	Systémy s transferem	133
4.3.3	Syntaktické struktury	135
4.3.4	Vztah mezi úlohami jazykové analýzy a syntézy	146
4.4	Systémy 2. generace: metody, nástroje a otevřené problémy	147
4.4.1	Základní principy analýzy	147
4.4.2	Valenční rámec — souhrnný popis „syntaktického chování“ slova	152
4.4.3	Zpracování viceslovných lexikálních jednotek	163
4.4.4	Problém anafor a katafor	166
4.4.5	Výrazy, které se nepřekládají	171
4.4.6	Problém výrazů nenalezených ve slovníku	172

4.4.7	Problém vlastních jmen	174
4.4.8	Závěrečná poznámka k problému indeterminismu syntaktické analýzy, homonymie a systematické vágnosti v přirozených jazycích	175
4.5	Potřeba věcných (mimojazykových, extralingvistických) znalostí při syntaktické analýze a transferu	176
4.6	Reálné možnosti, omezení a perspektivy aplikace počítačů v úloze překladu	180
4.6.1	Slabá místa, možné přínosy	180
4.6.2	Základní možnosti dělby práce mezi člověkem a počítačem na úloze překladu	183
4.6.3	Alternativní cesta — systémy s překladovou pamětí	199
4.6.4	Pravděpodobnostně řízená lexikální podpora překladatele — systémem <i>TransType</i>	204
4.6.5	Počítačem podporovaná tvorba překladových slovníků	205
5	Automatické získávání znalostí z textů	207
5.1	Základní problémy	207
5.1.1	Potřeba velmi přesného porozumění textu	207
5.1.2	Otázka optimální reprezentace znalostí	208
5.2	Obecný model analýzy a interpretace textu v přirozeném jazyce jako procesu manipulace se znalostmi	209
6	Automatizovaná korektura textů	213
6.1	Smysl, cíle a lingvistické úrovně	213
6.2	Pravopisná a „mechanická“ korektura	215
6.2.1	Rozpoznání chybně napsaných slov	215
6.2.2	Určení nejnadějnějších náhrad chybně napsaného slova	216
6.2.3	Může mít smysl opravovat některé detekované chyby plně automaticky?	217
6.2.4	Snadné doplnění pravopisné korektury — detekce „mechanických“ chyb	218
6.2.5	Podpůrné korekční funkce textových editorů	219
6.3	Gramatická korektura	221
6.3.1	Základní typy gramatických chyb	221
6.3.2	Aplikace vzorců popisujících syntaktické chyby	224
6.3.3	Obecněji použitelné přístupy a jejich omezení	226
6.4	Stylová korektura	229
7	Komunikace mezi člověkem a strojem v přirozeném jazyce	231
7.1	Konverzace jako cíl — program <i>ELIZA</i> a jeho následovníci	232
7.1.1	<i>ELIZA</i> a Turingův test	232
7.1.2	Algoritmus a datové struktury, v nichž spočívá „intelligence“ programu <i>ELIZA</i>	233
7.2	Komunikace s databázovým, resp. vyhledávacím systémem	236
7.2.1	Vyhledávání textových informačních zdrojů	236
7.2.2	Vyhledávání v relačních (a podobných) databázích	237
7.3	Zásady kooperativního dialogu	246
7.4	Syntéza vyjádření formalizovaných faktů v přirozeném jazyce	250

A	MOZAIKA — příklad metody automatického indexování maximálně signifikantními termíny z textu	265
A.1	Základní charakteristika, cíl a způsob použití metody <i>MOZAIKA</i>	265
A.2	Základní myšlenky a způsob fungování metody <i>MOZAIKA</i>	266
A.2.1	Morfologicko-lexikální analýza	266
A.2.2	Syntaktická analýza	273
A.3	Hodnocení, omezení, možnosti úprav	275
B	Přirozený jazyk z hlediska teorie formálních jazyků a automatů	277
B.1	Připomenutí základních pojmů a vět teorie formálních jazyků a automatů	277
B.2	Nejsou přirozené jazyky regulární?	278
B.3	Přirozený jazyk jako bezkontextový jazyk	281
C	Mírně zjednodušené schéma slovníku a lematizátoru použitého v českém pravopisném korektoru <i>PC Korektor</i>	285
D	Popis systému <i>AR2NL</i>	287
D.1	Základní funkční charakteristika systému	287
D.2	Hierarchická struktura lingvistických dat systému <i>AR2NL</i>	289
D.3	Jméno sledované entity a s ním spojené výrazy	289
D.4	Morfologické kategorie slov v systému <i>AR2NL</i>	290
D.5	Elementární fráze	290
D.5.1	Komentáře k tabulce D.2	291
D.6	Formulační vzorce	293
D.6.1	Základní komentáře k zápisu formulačních vzorců v XML	294
D.6.2	Adaptace formulačních vzorců na konkrétní počet literálů v antecedentu, resp. sukcedentu	295
D.6.3	Zjednodušený zápis formulačních vzorců	296
D.6.4	„Neterminály vyššího řádu“ ve formulačních vzorcích	297
D.6.5	Obhospodaření některých zvláštností slovosledu	298
D.7	Morfologický modul systému <i>AR2NL</i>	300
D.7.1	Soubor MN pro angličtinu	301
D.7.2	Soubor MN pro češtinu	302
D.7.3	Soubor MV pro češtinu	302
D.7.4	„Kvaziplurál“ českých sloves	303
D.8	Variabilita vyjadřování	303
D.9	Některé částečně rozpracované problémy	305
D.9.1	Možnost přeuspořádání konjunkce literálů	305
D.9.2	Možnost volby mezi spojkami „a“ a „ale“ při vyjadřování konjunkcí	306
D.9.3	Interference mezi výše popsanými jevy	307