Blanka Sedlačíková

Matematická lingvistika (4)

Učitel matematiky, Vol. 10 (2002), No. 4, 226–234

Persistent URL: http://dml.cz/dmlcz/150483

MATEMATICKÁ LINGVISTIKA (4)

BLANKA SEDLAČÍKOVÁ

Strojová lingvistika

Třetím dnes tradičně rozlišovaným odvětvím matematické lingvistiky (vedle lingvistiky kvantitativní a algebraické, které jsme si stručně představili v předcházejících částech tohoto článku) je **lingvistika strojová**. Vyvíjí se od konce padesátých let minulého století, a to zejména v souvislosti s rozmachem kybernetiky, kvantitativní a algebraické lingvistiky a jiných hraničních oborů. Vliv na její vznik měla také zvyšující se potřeba automatizace a mechanizace různých činností, tedy i práce s jazykem. Název strojová lingvistika nám říká, že se jedná o strojové zpracování jazyka, jež bylo prováděno zpočátku na jednoduchých děrnoštítkových strojích, později na složitých počítačích (odtud někdy označení **počítačová** nebo **komputační** lingvistika).

Postavení strojové lingvistiky v rámci ostatních oborů je následující: matematická lingvistika je tvořena 2 teoretickými obory, a to lingvistikou kvantitativní a lingvistikou algebraickou. Strojová lingvistika je potom jejich praktickou aplikací, proto se můžeme setkat i s termínem *aplikovaná matematická lingvistika.*

Počítače se využívají v celé řadě lingvistických úkolů, např. automatické generování vět podle určitých pravidel, různé statistické výzkumy apod. Mezi nejdůležitější problémy řešené strojovou lingvistikou patří *automatické uchovávání a vyhledávání informací, strojový překlad* a v současnosti zejména *korpusová lingvistika.*

Pro dnešní dobu je charakteristický neobyčejně rychlý nárůst informací všeho druhu. Většina vědeckých pracovníků není schopna sledovat záplavu článků a publikací vycházejících v jeho oboru. Je proto nezbytné, aby odborník prováděl jakýsi výběr. Nápomocny mu mohou být samozřejmě knihovní katalogy, v tisku uveřejňované soupisy literatury či résumé, recenze, informativní články

o těchto dílech apod. To ale nestačí, neboť takto zpracován je jen zlomek odborné literatury. Rozvoj výpočetní techniky podnítil myšlenku automatizovat **ukládání a vyhledávání informací**. V zásadě se jedná o tyto dva okruhy problémů: 1) vyhledání literatury, 2) informování o obsahu.

Vyhledávání literatury je úkol poměrně jednoduchý, neboť stroj pracuje pouze s termíny a jejich kombinacemi, a poradily si s ním už děrnoštítkové stroje. Jako příklad nám může posloužit klasifikace a indexování.

Cílem **klasifikace** je umožnit snadné nalezení literatury daného tématu. Dnes existuje několik stovek bibliografických systémů. Nejběžnější a mající dokonce mezinárodní platnost je *třídění desetinné*, které je založeno na principu tzv. silné hierarchie, tzn. že každý pojem má pouze jeden pojem nadřazený. Tak například v indexu 541.1 nám 5 může označovat matematiku a přírodní vědy, 54 chemii, 541 obecnou a teoretickou chemii a 541.1 fyzikální chemii. Fyzikální chemii můžeme rozdělit ještě do specializovanějších disciplín, kterým je pak přiřazeno další číslo od 0 do 9. Postupuje se od disciplín obecnějších k disciplínám speciálnějším a pro přehlednost se každá tři čísla oddělují tečkou. Nevýhodou tohoto systému je zařazování nových vědních oborů, neboť v rovině, do níž by obor patřil, bývá většinou všech deset čísel obsazeno. Proto vzniklo *třídění facetové*, založené na principu tzv. slabé hierarchie, v němž určitému pojmu může v rovině bezprostředně vyšší předcházet pojmů několik, což umožňuje snadné doplňování novými členy.

Úkolem **indexování** je co nejstručněji charakterizovat obsah nějakého dokumentu. Z práce se vyberou nejdůležitější pojmy a vhodně se zakódují, tzn. je jim přiřazen určitý index, který má zpravidla mnemotechnický charakter.

Automatické informování o obsahu odborných prací je záležitost mnohem náročnější. Kvalitní referát by měl být totiž jasný a srozumitelný, měl by být stručným obsahem práce, uvádět hlavní výsledky práce, její přínos a hodnocení. V zásadě mluvíme o dvou postupech automatického referování. První z nich využívá *statistiky*. Pro daný vědní obor se stanoví soubor nejdůležitějších

termínů a do referátu se automaticky přejímají ty věty, které obsahují nejvíce těchto termínů. Ze zkušenosti víme, že nejčastěji se vyskytující termíny nemusí být vždy nejvýznamnější z hlediska výstavby textu a také že důležité pasáže nemusí obsahovat žádný takový termín. Další postup, uplatňující hledisko *sémantické*, naráží ještě na větší problémy, neboť nelze pracovat bez velmi obtížné formalizace lexikálního významu.

Nejnáročnějším stupněm z postupů zpracovávajících informace pomocí počítače jsou **informační jazyky**, tzn. ucelené teoretické systémy sloužící k ukládání a vyhledávání informací. Snaží se řešit celou řadu již značně složitých úkolů. Například je žádoucí, aby stroj na základě souboru poznatků z odborné literatury daného oboru zodpověděl určitou otázku týkající se problematiky tohoto oboru tak, jak je v literatuře řešena, aby odpovídal na otázky kladené v přirozených jazycích, zpracovával literaturu psanou v různých jazycích, doplňoval informace, které nejsou v textu explicitně vyjádřeny apod. Dnes již existuje celá řada informačních jazyků lišících se obtížností, typem apod.

Strojovým překladem rozumíme převedení textu ze vstupního (výchozího) jazyka do jazyka výstupního (cílového) pomocí stroje (počítače). První pokusy byly provedeny v roce 1954 v USA a roku 1955 v SSSR. Počáteční nadšení, očekávající rychlé a levné překlady jakýchkoliv textů do nejrůznějších jazyků, bylo vystřídáno zhruba od poloviny šedesátých let do počátku let sedmdesátých minulého století značným rozčarováním. Překlady jednoho slova po druhém bez důkladnějšího jazykového rozboru nejsou možné. Pro strojový překlad je totiž nutné nejprve provést detailní **analýzu jazyka** výchozího ve všech rovinách jazykového systému a rovněž stejně detailní **syntézu jazyka** koncového a konečně sestrojit tzv. převodní jazyk (umělý jazyk, který by pomocí symbolů zachycoval významovou strukturu libovolné věty) či jinak umožnit automatické převádění z jednoho jazyka na druhý. Dnes víme, že kvalitní strojový překlad libovolného textu v zásadě není možný. Můžeme získat buď překlady málo kvalitní, které slouží pro hrubou orientaci v textu, nebo překlady kvalitní, ale pouze u textů

specializovaných (technických textů, manuálů apod.). Dobré výsledky lze také získat upuštěním od plné automatizace překladu, a to preeditací či posteditací uživatelem nebo tzv. interaktivním překladem, kdy ve sporných případech zasahuje člověk.

Velkým pomocníkem při tvorbě strojových překladů by se mohla stát **korpusová lingvistika**, která zaznamenává v současnosti obrovský rozvoj. Korpusovou lingvistikou rozumíme tu část počítačové lingvistiky, která se zabývá tvorbou a využitím jazykových korpusů, tzn. souborů jazykových dat. Tento soubor může mít podobu textů, které jsou zachyceny na papíře (nejčastěji v podobě excerpt), nebo může mít podobu elektronickou, tzn. že jazyková data jsou uložena v počítači. **Korpus** lze potom chápat jako rozsáhlý, vnitřně uspořádaný a ucelený soubor jazykových dat, která jsou elektronicky uložena, zpracována a přístupna. Ačkoliv se různých jazykových korpusů uchovaných na papíře užívalo velmi dávno (vzpomeňme na J. A. Komenského, kterému při požáru v Lešně shořel celý jeho rozsáhlý materiál na připravovaný latinsko-český a česko-latinský slovník), o skutečné korpusové lingvistice mluvíme až v souvislosti s rozvojem počítačové techniky.

Aby byl sestavovaný korpus dostatečně kvalitní, musí splňovat několik základních požadavků. Korpus by měl:

- být dostatečně *rozsáhlý*. Výsledky získané z malého množství dat mohou být zkreslené. Sledovaný jev nemusí být v malém korpusu zachycen vůbec nebo se může vyskytnout jen náhodně. Navíc by měl korpus reflektovat skutečný poměr jazykových jevů podstatných a jazykových jevů okrajových.

- obsahovat co nejvíce *variant* jazyka. Dříve se korpusy zpravidla omezovaly pouze na texty z oblasti umělecké literatury a publicistiky. Ve skutečnosti je použití jazyka daleko širší, proto by měl korpus obsahovat jak jazyk psaný (nejen styl umělecký a publicistický, ale i administrativní, odborný, texty soukromého rázu - korespondence, deníky apod.), tak jazyk mluvený (získaný od lidí různého věku, pohlaví, vzdělání, sociální úrovně apod.). Vedle spisovného jazyka by měl

být zachycen i jazyk hovorový a nářečí. Navíc by se tu kromě těchto synchronních složek měla vyskytovat část diachronní, tzn. záznam textů historických.

- zachycovat autentický jazyk, který odpovídá současnému úzu.

- být schopen neustálé aktualizace.

Splňuje-li korpus tato kritéria, je reprezentativním vzorkem daného jazyka – mluvíme o tzv. **národním jazykovém korpusu**. Ručně shromažďované korpusy budou naproti tomu vždy omezené co do počtu textů, variability, autenticity i aktualizace záznamů.

Při **sestavování počítačového korpusu** lze *ukládat texty* do paměti počítače trojím způsobem, a to:

1. v podobě *elektronické sazby*, která je dnes běžná při vydávání většiny knih, novin a časopisů a která se jeví jako způsob nejrychlejší a nejlevnější;

2. *skenováním* (pomocí OCR – Optical Character Recognition); 3

3. *ručním přepisem*, což je časově i finančně nejnáročnější způsob.

Takto získaná data jsou tzv. data vnější (hrubá) a je třeba je upravit a získat tak data vnitřní (upravená). Proto se jazykové soubory *čistí* od překlepů, tiskových chyb, nadbytečné grafiky apod.

Vyčištěné texty se *konvertují* do jednotného ASCII formátu (American Standard Code for Information Interchange), který zajišťuje standardizaci korpusů v mezinárodním měřítku.

Tato data se vyskytují v podobě lineárních řetězců a lze s nimi pracovat jen omezeně. S ohledem na další využití se proto provádí *značkování* korpusu (tagování, tagging; z angl. tag = visačka). Znamená to, že jednotlivým slovním tvarům je přidán index, který je blíže identifikuje. Značkování se provádí automaticky pomocí *morfologických analyzátorů*.

Pokud je slovním tvarům přiřazeno i jejich lemma (tj. základní podoba), pak mluvíme o tzv. *lemmatizátoru*. Pro zajímavost si uveďme, že morfologický lemmatizátor češtiny LEMMA (Ševeček, Osolsobě) pracuje s 1665 různými značkami (v angličtině je to nejvíce kolem 200 značek). Například slovnímu tvaru *matematik* by byla přiřazena značka k1gMnSc1, což znamená, že se jedná o slovní druh substantivum (k=1) rodu mužského životného (g=M), které se vyskytuje v singuláru (n=S) a prvním pádě (c=1).

Při automatickém značkování je z důvodů homonymie řadě slov přiřazeno značek více, proto je třeba provést tzv. *desambiguaci* (zjednoznačnění). Existuje řada metod, které provádí výběr tagu automaticky. Nejčastěji se užívá desambiguačních programů založených na statistických a pravděpodobnostních přístupech. Pro angličtinu se počet automaticky označkovaných slovních tvarů blíží 96 – 98%, v češtině to je přibližně 80%. Souvisí to s tím, že čeština je na rozdíl od angličtiny jazyk flektivní. V průměru připadnou v angličtině na 1 slovní tvar 2 tagy, zatímco v češtině je to tagů 6.

Z takto označkovaného korpusu je možno získávat seznamy slovních forem v jejich přirozeném kontextu – *konkordance*. V češtině se běžně používá formát KWIC (Key Word in Context), v němž lze při vyhledávání uplatnit různá hlediska – zobrazení výskytu jednotlivých slov, výběr podle charakteristiky slova. Hledané slovo se nám zobrazí v kontextu několika slov zleva i zprava (rozsah lze volit). Rejstřík slovních tvarů je možno třídit podle levého i pravého kontextu. Pomocí těchto nástrojů je například sestavování frekvenčních či retrográdních slovníků poměrně jednoduchou záležitostí, neboť velmi rychle a jednoduše můžeme získat různé statistické informace.

Ve světě dnes existují korpusy pro celou řadu jazyků (němčina, francouzština, angličtina, italština, španělština, maďarština, holandština aj.). První začaly vznikat na počátku šedesátých let. Největší počet jich je vytvořen pro angličtinu (přes 20). Za průkopníka korpusové lingvistiky je považován Randolph Quirk, který sestavil milionový *Survey of English Usage Corpus* britské angličtiny, který obsahuje polovinu psaného a polovinu mluveného

jazyka. První počítačový korpus *Brown Corpus of Written American English* byl vytvořen na Brown University v USA pod vedením W. N. Francise a H. Kučery v letech 1961 – 1964. Jeho rozsah je milion slovních tvarů. Svým složením jej kopíruje *Lancaster-Oslo/Bergen Corpus of British English* z roku 1970 sestavený pro britskou angličtinu. *London-Lund Corpus of Spoken English* z roku 1990 je složen z půl milionu slov mluvené angličtiny. Technika skenování byla poprvé použita u korpusu *Birmingham Collection of English Texts* a *Longman/Lancaster English Language Corpus* americké angličtiny. Největším korpusem je *Bank of English*, jehož rozsah by měl být 500 milionů slov.

K nejznámějším korpusům patří *British National Corpus*, který vznikl ve spolupráci lancasterské a oxfordské univerzity, nakladatelství Longman a Oxford a British Library roku 1991 a který obsahuje 100 milionů slov, z nichž 90 10% jazyk mluvený.

Vedle těchto korpusů zachycujících stav jazyka v daném časovém období, tj. korpusů synchronních, existují i korpusy diachronní (např. *Helsinki Corpus*). Většina korpusů je jednojazyčných, ale vznikají i vícejazyčné (např. anglicko-francouzský *Hansard*). Vzniká ale i celá řada specializovaných korpusů zaznamenávajících jen určitou část jazyka, například dialekty, díla jednoho autora, dětský jazyk, jazyk novin, právní předpisy aj. Obecný korpus může být pak složen z několika takových specializovaných subkorpusů.

I pro češtinu vzniká reprezentativní **Český národní korpus** (ČNK). Vůbec prvním korpusem češtiny byl manuálně zpracovaný korpus jazyka věcného stylu (VS) o rozsahu 540 000 slovních tvarů, který byl pod vedením Marie Těšitelové vytvářen v úseku matematické lingvistiky Ústavu pro jazyk český ČSAV od roku 1972. Celá řada pracovišť sestavovala pro své interní potřeby různé slovníkové databáze, jejichž využití bylo ale v důsledku neinformovanosti a nekoordinovanosti značně omezené.

Proto roku 1988 vzniká Iniciativní skupina pro přípravu počítačových korpusů, textů a slovníků, která si klade za cíl zajistit kompatibilní počítačové a programové vybavení na vysoké úrovni. Bohužel se tato snaha setkala s nezdarem. V roce 1991 je několika

odborníky z UK v Praze, MU v Brně a UP v Olomouci založena *Skupina pro počítačový fond češtiny*, jejímž cílem je vytvoření jakéhosi základního počítačového fondu češtiny – tzn. rozsáhlého textového korpusu a počítačových nástrojů určených k jeho zpracování. Roku 1992 začíná vznikat *Český národní korpus* jako grantový projekt *Čeština ve věku počítačů* (díky podpoře GA ČR) a na jeho budování se podílí odborníci z FF UK, MFF UK, FF MU, FI MU a ÚJČ ČAV. Roku 1994 je založen na FF UK *Ústav Českého národního korpusu*, který od podzimu 1995 práci na tvorbě ČNK koordinuje.

Cílem by mělo být vytvoření rozsáhlého, reprezentativního, elektronicky zpracovávaného souboru textů. Ten by měl posloužit zejména jako podklad pro sestavení nového českého výkladového slovníku a vytvoření nového frekvenčního slovníku češtiny. Korpus může sloužit jako zdroj materiálu pro lingvisty (frekvenční a statistické studie, ověřování lingvistických hypotéz, výzkum slovní zásoby apod.) i neligvisty (programátory, matematiky k ověřování nových automatických nástrojů pro výzkum jazyka, sociology, psychology, historiky, žurnalisty, učitele aj.). Výsledky získané korpusovou lingvistikou mohou být přínosné pro počítačovou lingvistiku, zejména strojový překlad.

V současnosti je ČNK složen z části *synchronní* a *diachronní*. Synchronní složku tvoří reprezentativní, vyvážený korpus současné psané češtiny SYN2000 o rozsahu 100 milionů slovních tvarů a korpus mluveného jazyka ORAL PMK. Plánuje se ještě vytvoření synchronní ho nářečního korpusu. Korpus SYN 2000 je sestaven z elektronických vydání některých našich periodik (Hospodářské noviny, Mladá fronta Dnes, Literární noviny, Vesmír, Vlasta, Divadelní revue aj.), produkce některých nakladatelství (např. Nakladatelství Lidových novin, Mladá fronta, Trizonia, Atlantis) a zařazeny jsou i všechny nově přijímané zákony České republiky. Jeho součástí je subkorpus PUBLIC, který obsahuje 20 milionů slov (procentuální zastoupení je stejné jako v SYN2000) a je přístupný po internetu. Část mluvená ORAL je zatím reprezentována pouze Pražským mluveným korpusem (PMK), ale vzniká subkorpus brněnských mluvčích (očekávaný rozsah asi 500 tisíc

slovních tvarů; zatím existuje ve formě přepisu nahrávek a má už částečně označkovanou podobu, z toho asi 30 000 slov je označkováno úplně. Část korpusu tvořená kvalitními nahrávkami byla převedena do digitalizované podoby na CD). Diachronní část tvoří korpus historické češtiny DIAKORP a plánuje se i diachronní nářeční korpus. Mimo to byl koncem roku 1997 v Brně (spoluprací mezi FF MU a FI MU) vytvořen označkovaný synchronní korpus DESAM, který obsahuje 1 026 733 slovních tvarů. Nově vzniká na FF MU i subkorpus soukromé korespondence. Podrobnější informace o ČNK můžete najít a práci s korpusem PUBLIC (pomocí korpusového manažeru CQP) si vyzkoušet na internetové adrese `http://ucnk.ff.cuni.cz/menu.html`.

Literatura

[1] Čermák, Fr., Blatná, R., *Manuál lexikografie*, H& H, Jinočany, 1995

[2] Černý, J., *Dějiny lingvistiky*, Votobia, Olomouc, 1996

[3] Hlaváčová, D., *Korpus mluvené češtiny*, diplomová práce, Brno, 1998

[4] Hvězdová, B., *Tvoření adverbií paradigmaticky odvozených od adjektiv na materiálu ČNK*, diplomová práce, Brno, 1999

[5] Sgall, P., kol., *Cesty moderní jazykovědy*, Orbis, Praha, 1964

[6] Těšitelová, M., *Kvantitativní lingvistika*, SPN, Praha, 1987

[7] Vašák, P., *Matematika, exaktnost a literatura*, Československý spisovatel, Praha, 1986

Mgr. Blanka Sedlačíková
doktorandka Katedry matematiky PřF MU
Janáčkovo nám. 2a, 662 95 Brno
e-mail: hvezdova@math.muni.cz