Počítačová lingvistika ve vztahu k informatice

Jarmila Panevová a kolektiv*

1 Úvod

Vývoj počítačových technologií a počítačového zpracování jazykových dat spolu s převratnými změnami v obou složkách jsou spolu zcela neodmyslitelně spjaty. Jejich vzájemnou souvislost i podmíněnost nových možností automatického zpracování přirozeného jazyka v kontextu nových informačních technologií se tu pokusíme ukázat z hlediska dneška ([9]). Na začátku se letmo podívejme na nedávnou minulost. Obor u nás tehdy nazývaný matematická lingvistika se konstituoval v 50. letech. Podle používaných přesných metod se tehdy rozvíjel ve dvou směrech: jako lingvistika algebraická a statistická. K prvním experimentům s jazykovými daty algoritmicky zachycenými (pro účely strojového překladu, automatického vyhledávání informací v textu, komunikace s databázemi v přirozeném jazyce, při vývoji expertních systémů ap.) se používaly počítače starších generací (SAPO, EPOS, MINSK, ROBOTRON) s off-line přístupem. Připomínáme to proto, abychom ukázali nejen nevídané možnosti dané technickým pokrokem v posledních 50 letech, ale i jisté teoretické zázemí a zkušenosti s formalizovaným popisem jazyka, které se tu nahromadily. Dynamika vývoje v obou složkách se odráží i v terminologii: dnes se mluví o formální (popř. o teoretické) a o komputační a korpusové lingvistice.

Razantní nástup výpočetní techniky, kterého jsme svědky v posledním desetiletí, s sebou přinesl nové možnosti i v lingvistice. Vydávání novin, knih a časopisů je dnes nemyslitelné bez elektronické podoby vstupních textů. To na jedné straně klade nové požadavky na formulaci poznatků o jazyce (vyvolané např. potřebami kódování jednotlivých grafématických systémů, počítačových modulů pro "korekci" pravopisu, tzv. spell-checkerů, pravidel pro dělení slov v textových editorech ap.), na druhé straně to pro počítačovou a korpusovou lingvistiku skýtá obrovské báze slov a textů. Sbírky textů i mluveného slova v řádech stovek miliónů výskytů slov, jež dříve prostě nebylo kam uložit, dnes slouží jazykovědcům při budování slovníků a dokládání nejrůznějších jazykových

*Prof. PhDr. Jarmila Panevová, DrSc., Ústav formální a aplikované lingvistiky (ÚFAL) MFF UK, Mgr. Jan Cuřín, Mgr. Martin Čmejrek, Mgr. Nino Peterek, Mgr. Daniel Zeman, studenti doktorského studia v ÚFAL; Mgr. Barbora Hladká Ph.D., Mgr. Kiril Ribarov, odb. pracovníci Laboratoře pro zpracování jazykových dat při ÚFAL MFF UK; RNDr. Vladislav Kuboň, odb. asistent ÚFAL MFF UK. Práce byla podporována grantovými projekty VS 96151 a GAČR 405/96/K214.

jevů. Díky dostupnosti velkého množství dat (korpusů textů, korpusů mluveného slova) se začala rozvíjet korpusová lingvistika.

Jeden ze spojovacích článků korpusové lingvistiky a lingvistiky komputační představují anotované korpusy. Chápeme-li korpus jako soubor "surových", elektronicky uložených textů či promluv s jednou dobře definovanou strukturou, pak anotovaný korpus je korpusem obohaceným o ručně přiřazené informace nejrůznějšího charakteru. Prudký rozvoj zaznamenávají metody počítačového zpracování přirozeného jazyka založené na informacích, které anotovaný korpus nabízí. Jde tedy o strojové učení, o snahu modelovat jazykový systém na základě příkladů "ze života", tj. z existujících textů a promluv.

Zkušenosti lingvistů ukazují, že popsat jazyk souborem přesných pravidel je mimořádně obtížné (spekuluje se i o tom, zda je to vůbec možné - to je však diskuse spíše filozofická, která vede až k diskusím o možnostech umělé inteligence vůbec), pokud mají být zachycena všechna zákoutí jazyka, všechny výjimky. Praktické zkušenosti "korpusových informatiků" ukazují, že analýzou velkého množství dat lze někdy odhalit zákonitosti, které jazykovědce-člověka nenapadnou.

Moderní směry počítačové lingvistiky (a tím i lingvistiky korpusové) zaznamenávají velký, avšak různorodý vývoj. Pokusíme se je klasifikovat podle dvou hledisek: (1) na základě typu modelů použitých k popisu přirozeného jazyka; (2) na základě způsobu použití (předem) zpracovaných lingvistických znalostí.

(1)
- (a) formální lingvistický teoretický popis přirozeného jazyka za použití formálních gramatik (různých modifikací bezkontextových, kontextových a závislostních gramatik), automatů, logických a algebraických metod
- (b) modely založené na pravidlech
- (c) stochastické modely
- (d) modely používající neuronové sítě, genetické algoritmy
- (e) různé hybridní přístupy, které nejrůznějšími způsoby kombinují nějakou podmnožinu z výše jmenovaných přístupů

(2)
- (a) algoritmus kóduje jazykovou znalost (tj. lingvisticky zjištěná a popsaná pravidla)
- (b) algoritmus zakódovaná pravidla nemá, má však definovaný postup, jak tato pravidla určovat, včetně toho, jak je používat

Přístup (2b) je závislý na korpusové lingvistice. Pravidla extrahuje z dříve (lingvisticky) zpracovaných textů. Algoritmus má být navržen tak, aby mohl tato pravidla dostatečně zobecnit a aplikovat i pro texty neznámé. U algoritmů (2b) se lingvistické aspekty promítají jak nepřímo přes existenci anotovaných jazykových korpusů, tak i přímo, při vytváření jazykového modelu (1c), pří výběru typu pravidel (1b), při návrhu sítě (1d) a ve všech případech spadajících pod (2b) včetně selekce dat k učení.

2 Aplikace korpusového statistického modelování

V rámci korpusového modelovaní přirozeného jazyka[1] (viz bod 2b) potřebujeme více než jen megabajty elektronicky uložených textů. Texty je nutné projít a ručně v nich vyznačit to, co chceme u dalších najít automaticky. Touto namáhavou prací se platí za možnost rychlého vyhledávání nejrůznějších statistik, které se týkají zpracovávaného jazyka. Místo aby lingvista trávil měsíce a roky nad formulací pravidel, která daný jazyk co nejpřesněji popíšou v rámci konkrétní aplikace, věnuje svůj čas doplnění analyzovaných údajů do reálných dat a s vědomím jistého rizika spoléhá, že pravidla bude schopen odhalit počítač.

Pro potřeby korpusového modelování je ručně zpracovaný anotovaný korpus v nejjednodušším případě rozdělen na dvě části - trénovací korpus (90% celkového objemu) a testovací korpus (10% celkového objemu). Trénovací korpus reprezentuje materiál pro trénování (učení se) daných skutečností. Naopak testovací korpus je materiálem pro vyhodnocení kvality natrénovaného modelu (jak jsme se příslušným modelem přiblížili ke skutečnému jazyku). Některé metody vyžadují vyčlenit z anotovaného korpusu tzv. ladicí korpus, který slouží pro ladění parametrů zvoleného algoritmu.

Přístupy založené na statistickém modelování patří mezi nejrozšířenější přístupy založené na korpusech. Uvedeme několik příkladů, kdy mohou postupy známé ze statistiky a pravděpodobnosti pomoci s úkolem pro počítač zdánlivě neřešitelným. V prvním půjde o zpracování mluvené řeči (2.1), v dalším budeme požadovat, aby počítač určil morfologický tvar slova (2.2), ve třetím pak stavbu celé věty (2.3), v posledním (2.4) se pokusíme popsat uplatnění těchto metod i na strojový překlad, dříve klasickou doménu aplikací algebraické (strukturně orientované) lingvistiky. V rámci daných aplikací se pokusíme upřesnit údaje uvedené v předchozím oddíle a zmíníme se o dosavadních zkušenostech získaných statistickým modelováním češtiny.

2.1 Rozpoznávání mluvené řeči

Podaří-li se zvládnout automatické zpracování mluvené podoby jazyka, získají lidé klíč k téměř bezbariérové komunikaci s počítačem. V současné době je jednou z nejúspěšnějších metod zpracování mluvené řeči modelování fonetických jednotek prostřednictvím skrytých Markovových modelů - generujících konečněstavových stochastických automatů (dále jen HMM). Zpracováváme-li akustický signál promluvy, předpokládáme, že je výsledkem zakódování symbolického zápisu. Při analýze tohoto signálu se snažíme odhadnout jeho textovou podobu. Vlastní akustický signál přijímáme jako posloupnost naměřených hodnot (vektorů příznaků) tvořících časovou řadu. Abychom ze signálu získali symbolický zápis prom-

[1] V r. 1994 vznikl na Filozofické fakultě Univerzity Karlovy v Praze Ústav českého národního korpusu (vedený prof. Františkem Čermákem), v němž se shromažďují a zpracovávají české texty v elektronické podobě. Jsou to texty psané i mluvené, jsou unifikovány, "čištěny" a zpracovávány spolu s vývojem a zdokonalováním "korpusového manažeru", s jehož pomocí v nich lze pohodlně vyhledávat data tříděná podle potřeb konkrétního uživatele. Český národní korpus tak vytváří nebývale rozsáhlou datovou základnu umožňující mj. sledovat i aktuální úzus našeho národního jazyka.

luvy, musíme znát kódovací funkci, která je v našem případě modelována pomocí HMM. Je-li naším cílem rozpoznávání izolovaných slov, je situace relativně jednoduchá, protože nemusíme hledat hranice slov. Máme-li několik modelů M_1, M_2, ...,M_m reprezentujících m slov slovníku a časovou řadu hodnot $O = o_1$, o_2, ..., o_T patřící nějakému slovu, můžeme zjistit, který z modelů by danou časovou řadu vygeneroval s největší pravděpodobností. Tyto pravděpodobnosti se počítají procházením modelů od počátečního stavu, přičemž se snažíme projít všechny cesty automatu o délce T. Pro každou z těchto cest končící v některém z koncových stavů automatu spočítáme pravděpodobnost. Ta je rovna součinu pravděpodobností všech přechodů cesty a pravděpodobnosti, s jakou generuje každý k-tý stav cesty k-tou hodnotu časové řady, což je zjistitelné z funkce rozložení pravděpodobnosti k-tého stavu. Součet pravděpodobností těchto cest je roven pravděpodobnosti, s jakou model danou řadu vygeneroval. Výpočet provedeme na všech modelech . Model, který s největší pravděpodobností vygeneroval danou časovou řadu O, zároveň reprezentuje i nejpravděpodobnější vyslovené (analyzované) slovo.

Dostatečnou aproximací předchozího postupu je hledání maximálně pravděpodobné cesty automatem, která by generovala danou časovou řadu. Maximálně pravděpodobnou cestu efektivně nalezne Viterbiho algoritmus. Parametry HMM pro každé slovo lze zcela automaticky natrénovat jeho ukázkovými promluvami a Baum-Welchovým algoritmem.

Pro rozsáhlé slovníky je však nevhodné počítat model pro každé slovo. Modelování fonémů, případně kontextově závislých fonémů (trifonémů), umožní pokrýt jazyk mnohem menším množstvím modelů. K rozpoznávání je pak zapotřebí fonetický slovník tvarů rozpoznávaného jazyka. Před rozpoznáváním se vybuduje velká HMM síť, která pro každé slovo zřetězí modely fonémů na základě fonetického slovníku. Všem těmto řetězcům se předřadí startovní negenerující uzel a zakončí se společným negenerujícím koncovým uzlem, z něhož vede hrana opět do startovního uzlu. Je-li takovéto síti předána časová řada vektorů příznaků reprezentující nahranou větu, pak je maximálně pravděpodobná cesta zjištěná Viterbiho algoritmem zároveň i fonetickým přepisem celé vyslovené věty. Každý z průchodů koncovým stavem je i vyjádřením hranice slov ve větě.

Pro rozpoznávání češtiny používáme rozšířený program HTK, tj. softwarový nástroj pomáhající při vytváření automatického rozpoznávače, a to od přípravy trénovacích dat až po analýzu úspěšnosti rozpoznávače. Aby byl tento program pro češtinu použitelný, museli jsme zvolit vhodný fonetický přepis, najít reprezentativní trénovací texty a řešit problémy spojené s volným slovosledem a flexivností jazyka. Těžiště dosavadní práce bylo spíše v hledání grafického zápisu mluvených projevů a reprezentativních trénovacích textů. Volný slovosled i flexi jsme řešili omezením jazykové domény definované slovníkem menšího rozsahu a jednoduchou gramatikou. Úspěšnost rozpoznávání pro češtinu se v současnosti pohybuje okolo 80 % správně rozpoznaných slov ([2]). Je přitom používáno slovníku s 63 tisíci slov a kontextově závislých akustických modelů natrénovaných na 12 hodinách mluvené řeči od 47 mluvčích. Současné úsilí je zaměřeno na zlepšování jazykových modelů, analýzu intonace a rozšíření kolekce nahrávek spontánní řeči nejlépe reprezentujících mluvenou řeč.

2.2 Morfologické značkování

Chceme-li nejen zjistit základní podobu slova z textu (jeho lema), ale i jeho morfologické údaje (a takové jsou zpravidla požadavky kvalifikovaného uživatele textového korpusu jako zdroje dat), musíme přiřadit jednotlivým výskytům slovních forem jednoznačnou morfologickou informaci, která je převážně z technických důvodů reprezentována jako morfologická značka. V rámci sestavení repertoáru morfologických značek je možné se pohybovat od obecných značek pokrývajících pouze informaci o slovním druhu (podstatná jména, přídavná jména, slovesa atd.) až po úplné značky zachycující úplnou morfologickou strukturu daného jazyka (např. pro podstatná jména rod, číslo a pád). Výstupem automatické morfologické analýzy je pro slovní formu množina všech možných značek, které pro danou slovní formu připadají v úvahu (pro slovní tvar *pře* to bude jistý tvar podstatného jména a tvar slovesa, pro tvar *při* to bude opět tvar téhož podstatného jména, imperativ slovesa a předložka). Procedura automatického morfologického značkování pak vybírá z této množiny právě jednu značku, o níž se na základě kontextové informace předpokládá, že je správná. Slovo *stav* morfologická analýza vyhodnotí jako sloveso a zároveň jako podstatné jméno. Teprve v rámci dané věty, popř. odstavce (tedy na základě kontextu) je jak člověk, tak automat schopen tuto nejednoznačnost vyřešit.

Pro řešení této úlohy je třeba především mít k dispozici korpus opatřený manuálně přiřazenými morfologickými značkami. Na cestě k automatickému morfologickému značkování jsme pro češtinu použili dva stochastické modely - Markovovy modely ([12]) a exponenciální model ([7]). V rámci experimentů morfologického značkování českých textů zkoumáme vlivy jednotlivých faktorů (velikost trénovacích dat, specifičnost morfologických značek, předzpracování textu morfologickou analýzou, atd.) na celkovou úspěšnost značkovací procedury. Úspěšnost značkovací procedury je možné hodnotit absolutně i relativně. Při absolutním hodnocení nás zajímá pouze experiment s nejvyšším procentem úspěšnosti. Naopak při relativním hodnocení věnujeme pozornost nejen procentu úspěšnosti, ale také nás zajímají hodnoty jednotlivých parametrů experimentu. V případě značkování pomocí obecných značek jsme sice pro češtinu získali nejlepší výsledky, ale např. pro syntaktické zpracování textu automaticky označkovaného obecnými značkami se pouhá informace o slovním druhu jeví jako výrazně nedostatečná. Markovovými modely i exponenciálním modelem češtiny jsme se přiblížili k hladině 94% úspěšnosti.

Morfologické značkování nelze ovšem chápat jako finální aplikaci, je však u flektivního jazyka nezbytným předpokladem pro uplatnění dalších kroků automatické analýzy. Morfologicky označkovaný text slouží pro kvantitativní výzkum, pro vyhledávání v korpusu nebo je přímo vstupem dalších procedur automatického zpracování přirozeného jazyka.

2.3 Syntaktická analýza

Známe-li nyní ke každému slovu ve větě jeho morfologickou značku, můžeme se pokusit o nalezení vztahů mezi jednotlivými slovy. Mnozí z nás si jistě pamatují ze školy větný rozbor, v němž jsme měli za úkol uspořádat větné členy do závislostní struktury, jíž odpovídá stromový graf. Je zřejmé, že znalost vztahů mezi řídícími a závislými členy je klíčová pro porozumění významu věty, a právě tak automatické nalezení těchto vztahů – automatická syntaktická analýza – je klíčem k počítačovému překladu z jednoho jazyka do druhého, k nalezení chyb v textu (například špatné *i/y* ve větě *"Muži přišly."*) i k dalším aplikacím, jež takovou hloubku analýzy vyžadují.

I když můžeme pouze spekulovat, co se odehrává v mozku každého z nás při takové analýze, lze tyto spekulace opřít o různá pozorování. Víme například, že podmětem věty je velmi často podstatné jméno v prvním pádě, přísudkem zase sloveso. Přídavná jména často rozvíjejí podstatná jména, s nimiž se shodují v rodě, čísle a pádě. Takových a podobných omezujících podmínek lze vymyslet desítky a stovky, od jednoduchých, jako jsou tyto zde uvedené, po velice komplikované. Třeba z věty "*Pes naší babičky nekouše*" usoudíme, že jmenná fráze ve druhém pádě rozvíjí podstatné jméno (to je její prototypické postavení), ale takové pravidlo nebude fungovat absolutně, např. nefunguje ve větách *"Ten pes je naší babičky*, *Svou poznámkou se silně dotkl naší babičky*". Korpusy nám nabízejí možnost projít věty, které už někdo pronesl či napsal, a vyřešit tak současně dva problémy: zaprvé zjistit, jaké závislosti vůbec existují, a za druhé jim přiřadit váhy, pravděpodobnosti jejich výskytu. Vybudujeme tak vlastně statistický model syntaktické struktury věty. Abychom mohli větnou skladbu modelovat, potřebujeme zvláštní anotovaný korpus, kterému se anglicky říká treebank, tedy něco jako zásobárna stromů, tj. grafů, které reprezentují syntaktickou strukturu vět. Uveďme si pro ilustraci několik takových modelů ([5], [8]). Ten nejzákladnější je založen na četnostech výskytů dvojic slov; "učíme" se z treebanku, a můžeme tedy evidovat dvojice slov, která na sobě závisí. Tváří v tvář nerozebranému textu pak zalovíme v tabulce a postavíme takový strom, aby závislosti, z nichž se skládá, měly co možná nejvyšší pravděpodobnost. Protože však různých slovních tvarů jsou statisíce a různých dvojic tedy potenciálně 10^{10}, hrozí nám, že treebank nikdy nebude dost velký, abychom alespoň jednou nalezli každou závislost mezi dvěma konkrétními slovy, která přichází v úvahu, a počítač nikdy nebude mít dost paměti, aby všechny tyto závislosti uměl uložit a účinně zpracovat. Proto můžeme využít již hotových morfologických značek a zjišťovat závislosti mezi nimi. Na světě je druhý model, který sice nedosahuje přesnosti prvního, ale dokáže kupříkladu vyjádřit shodu podmětu s přísudkem či přívlastku s rozvitým podstatným jménem. Připadá-li nám to jako příliš velké zjednodušení, můžeme evidovat dvojice řídící slovo - závislá značka a modelovat tak například slovesné vazby (na slovese *vidět* musí "záviset" předmět ve čtvrtém pádě, což neplatí např. o slovesu *věřit*). Další model může sledovat slovosled, např.: leží závislý člen těsně vedle řídícího, nebo někde dál ve větě? Takto můžeme pokračovat. Jak dobře se dokážeme přiblížit té správné větné struktuře, závisí nejen na velikosti treebanku, ale také na tom, jaké statistiky zvolíme a jakým způsobem je zkombinujeme.

2.4 Statistický strojový překlad

K finálním aplikacím statistického zpracování přirozeného jazyka patří také strojový překlad - jeden z nejsložitějších a nejkomplexnějších úkolů počítačové lingvistiky. V roce 1965, po kritickém zhodnocení tehdejších výsledků v oblasti strojového překladu americkou poradní komisí ALPAC, došlo ve Spojených státech k víceméně úplnému zastavení výzkumu v této oblasti na téměř třicet let. V Evropě se strojový překlad ubíral cestou ručního vytváření syntaktických pravidel a využívání překladových slovníků obohacených o sémantické informace. V posledních letech došlo k zvýšení zájmu o strojový překlad a k řešení tohoto problému se začínají využívat i metody matematické statistiky. Výhodou statistického přístupu je zejména obecnost, která umožňuje relativně snadno použít metody vyvinuté pro určitou jazykovou dvojici i pro další jazykové páry.

Nezbytným předpokladem použití statistických metod pro strojový překlad je existence paralelního korpusu. Paralelní korpus je tvořen texty ve dvou nebo více jazycích, tyto texty si svým obsahem a strukturou (pořadím odstavců a vět) odpovídají. Klasicky je jeden text manuálním (intelektuálním) překladem druhého. V ideálním paralelním korpusu jsou kromě jednoznačného párovaní odstavců a vět spárována i jednotlivá slova či sousloví. Takový korpus však bývá k dispozici zřídkakdy, většinou existuje pouze jednoznačné párování mezi celými dokumenty, a párování odstavců a vět je nutné provést dodatečně, zpravidla automaticky. K automatickému párování odstavců a vět se využívá jednoduchá statistická metoda založená na porovnávání délek příslušných pasáží, jejíž výstup může být zpřesněn zohledněním shodných nebo podobných částí textu v obou jazycích (číslic, jmen či přejatých slov) nebo informací z překladových slovníků. Paralelní korpus s vyznačenými paralelními větami slouží jako trénovací data pro překladový model. Minimální velikost korpusu použitelná pro statistický překlad je kolem 50 tisíc párů vět.

Překladový model je založen na pravděpodobnostním překladovém slovníku, tedy slovníku, ve kterém je jednotlivým překladům přiřazena pravděpodobnost, s jakou odpovídají příslušnému heslu. Parametry tohoto modelu (pravděpodobnosti párů heslo - překlad) jsou vypočítány iterativním EM-algoritmem: Na počátku se vytvoří slovník, v němž jsou u každého hesla uvedena všechna slova, se kterými se vyskytlo ve větném páru. Všechny překlady daného hesla mají zatím stejnou pravděpodobnost; zatím nevíme, které z nich jsou správné. Spočítáme-li pro každý možný překlad částečný součet pravděpodobností, do něhož zahrneme pro každou dvojici vět v korpusu pravděpodobnost, že se v dané větě pro příslušné heslo použil právě onen překlad, získáme pro různé překlady různé hodnoty. Pravděpodobnosti překladů každého hesla potom rozdělíme v poměru těchto částečných součtů. Protože pravděpodobnost, že se při překladu věty použil právě ten či onen překlad hesla, závisí na pravděpodobnosti překladu hesla, dostaneme v další iteraci jiné částečné součty. Celý systém však po několika iteracích konverguje k hledanému řešení. Odfiltrováním málo pravděpodobných překladů můžeme získat překladový slovník pro lidské uživatele.

Princip trénování základního modelu používáme i při vytváření modelu rozšířeného o další množiny parametrů. Navíc bereme v úvahu tabulku změn pozic slov ve větě a tzv. tabulku "plodnosti" slova, která obsahuje pravděpodobnosti, s jakými z jednoho slova ve zdrojovém jazyce vznikne n slov v jazyce cílovém, nebo naopak, že toto slovo nebude mít ve druhém jazyce ekvivalent. Je možné stanovit i globální koeficient zkrácení či prodloužení vět při překladu.

Při vlastním statistickém strojovém překladu pracujeme kromě překladového modelu i s modelem jazykovým, který by měl zaručit, že sestavená věta je v cílovém jazyce přípustná. Takovým jazykovým modelem může být například jednoduchý trigramový model (model založený na tabulce relativní četnosti trojic po sobě následujících slov) sestavený na základě jednojazyčných textů z korpusů cílového jazyka, kterých je k dispozici více než textů paralelních. Překlad věty probíhá tak, že k překládané větě hledáme větu v cílovém jazyce, která má z hlediska překladového a jazykového modelu nejvyšší pravděpodobnost.

Data z paralelního anglicko-českého korpusu (rozděleného na část beletristickou – převážně články z časopisu Reader's Digest, obsahující 50 tisíc paralelních vět, a odbornou – překlady manuálů, lokalizace operačních systémů o délce 650 tisíc paralelních vět) byla použita při automatické extrakci česko-anglických slovníků z paralelních textů ([3]). Kvalita slovníků vytvořených na základě dat z odborné části korpusu byla srovnatelná s výsledky extrakce slovníků z velkého čínsko-anglického korpusu HKUST. V roce 1999 na letním semináři v Baltimore v USA ([14]) byly provedeny první experimenty s česko-anglickým statistickým strojovým překladem. Překvapivě dobrých výsledků bylo dosaženo při překladech vět z testovacího vzorku odborné části korpusu (30% vět automaticky přeložených z češtiny odpovídalo anglickému originálu). Při experimentech s česko-anglickým překladem beletristické části korpusu, s výrazně menším objemem trénovacích dat, se ukázalo být užitečným předzpracování české části textů lematizací. V průběhu jednoho měsíce byly implementovány obecné nástroje pro sestavování překladových modelů průběžně testované na překladech z francouzštiny do angličtiny a z češtiny do angličtiny.

3 Metody formální strukturně orientované

V odd. 2 jsme se věnovali aplikacím, které jsou založeny na stochastických metodách. Ty patří v současné době k rozšířeným směrům; jsou na jedné straně (statisticky) nejpřesnější, na druhé straně patří jejich užití k nejsnadnějším. Ukázali jsme ovšem i to, že tyto metody potřebují "trénovací" data, tj. data obsahující příslušné jazykové údaje (fonetické, morfologické, syntaktické). Čím kvalitněji jsou tato data připravena a čím je jich více, tím lepší výsledky lze očekávat od jazykového modelu, který "napodobováním" takto zpracovaných údajů vznikne. Plyne z toho ovšem i to, že intelektuální jazykovou analýzu nelze ani tady ničím nahradit. Zmínili jsme se o tom v odd. 1 a 2.3.

Jednou z oblastí počítačové lingvistiky, kde se statistické metody prakticky nedají uplatnit, je oblast zjišťování, klasifikace a opravování gramatických chyb v textech. Na první pohled by se mohlo zdát, že to, co jsou schopni se naučit žáci nižšího stupně základní školy, musí být lingvistickými metodami také zvládnutelné. To ovšem není pravda, protože problém gramatické správnosti vět přirozeného jazyka je velmi těsně spojen nejen s tvaroslovím či skladbou, ale závisí do značné míry také na významu a smyslu věty, což jsou oblasti pro algoritmický popis velmi obtížné.

Jádro problému kontroly gramatické správnosti si můžeme ilustrovat na jednoduchém příkladu české věty *"Koťata chytaly myši."* Každá učitelka češtiny jistě žákům opraví chybu ve shodě podmětu s přísudkem (*koťata chyta**la** myši*), protože je přece jasné, že podmětem věty jsou *koťata*. Z hlediska čistě skladebného ovšem nemá pravdu, protože podmětem mohou být i *myši*, a pak je věta zcela správně. V tomto případě tedy rozhodnutí o správnosti či nesprávnosti závisí na porozumění smyslu věty a hlavně na znalosti světa okolo nás, ve kterém přece jenom jsou myši chytány koťaty a nikoli naopak. Takovou hloubku znalosti světa ovšem nedokážeme počítačem zachytit nejen nyní, ale ani v blízké budoucnosti.

Vezmeme-li v úvahu výše uvedené příklady, je jasné, že na automatickou kontrolu gramatiky nesmíme mít maximalistické nároky. Pokud se ovšem spokojíme s tím, že automatická kontrola je s to odhalit pouze některé chyby, můžeme poměrně úspěšně uplatnit metody automatické gramatické korekce pomocí formální gramatiky. To ukázala například úspěšná realizace experimentálního modulu gramatického korektoru, která byla vyvinuta na MFF UK ([13]). Na jazykových pravidlech založená metoda je schopna odhalovat některé chyby ve shodě, v interpunkci i řadu dalších chyb. Výše uvedený příklad s koťaty naše metoda přijme jako gramaticky správnou českou větu.

4 Textový korpus – zdroj poznání o jazyku

Úlohu jazykových korpusů jsme už v odd. 1 - 3 dostatečně zdůraznili. Podívejme se nyní blíže na stavbu jednoho korpusu z mnoha, totiž Pražského závislostního korpusu ([10], dále PDT).

Texty vybrané z Českého národního korpusu vypracovaného na FF UK jsou značkovány na třech úrovních, z části manuálně, z části automaticky. Ke každé větě se tak pořizuje anotace na třech různých úrovních. Na první úrovni jsou slovům v textu přiděleny morfologické značky obsahující informaci o tvaru slova (rod, číslo a pád u jmen, osoba, číslo, čas atd. u sloves, srov. odd. 2.2 výše). Na druhé úrovni, nazývané analytická rovina, je z řetězce slovních tvarů automatickou procedurou a posléze její manuální modifikací vytvořena závislostní struktura odpovídající stavbě věty ([1], [4], [6]). Prostředky využívané na analytické rovině nejsou dostatečně jemné pro rozlišení všech významových distinkcí ve větné stavbě (příslovečná určení nejsou dále sémanticky tříděna na způsobová, lokální, temporální, kauzální atd., pomocná slova jsou tu reprezentována jako samostatné uzly stejně jako slova plnovýznamová, vypuštěná slova nejsou do stromu doplňována ap.). Analytická rovina slouží jako předstupeň pro vlastní "hloubkovou" reprezentaci zjednoznačněného významu věty, pro niž užíváme termínu "tektogramatická" stromová struktura ([11], dále TGTS). Tato reprezentace obsahuje jako své uzly již jen plnovýznamová slova (až na malé výjimky vynucené potřebou zachytit vícerozměrnou stavbu věty s vrstvami souřadných a aponovaných struktur v dvojrozměrném grafu). Uzly jsou spojeny hranami odpovídajícími typům syntaktických vztahů (determinace, koordinace, apozice, vsouvání). Typ vztahu mezi uzlem mateřským a dceřiným se označuje (ve velké většině případů manuálně) v hodnotě funktoru dceřiného uzlu. Členy elidované v povrchové stavbě věty se na TGTS doplňují, zachytí se u nich, zda jde o elipsu příležitostnou (textovou), nebo gramatikalizovanou. Ke každému uzlu se připojí i základní informace o jeho postavení v tématu (mezi "známými" prvky informační struktury věty) či rématu (mezi "novými" prvky) výpovědi.

Na analytické rovině bylo zatím označkováno cca 100 tisíc českých vět, které vstupují do fáze tektogramatického značkování; v této etapě jsou věty napřed automaticky předzpracovány (stromová struktura je "prořezávána"), to znamená, že jsou likvidovány pomocné uzly, ale uchovávána sémantická informace, která je v nich obsažena, dále jsou přidělovány některé funkce, které jsou jednoznačné, a poté anotátoři doplňují ty informace o tektogramatické struktuře, na něž zatím automatická procedura nestačí. V budoucnu podíl automatické procedury na tektogramatickém anotování poroste.

Takto označkovaný korpus je nejen podkladem pro "trénování" fonetických, morfologických, syntaktických a překladových procedur, ale je i neobyčejně cennou, kvalitativně i kvantitativně novou základnou pro jazykovědný výzkum. Z anotací lze snadno automaticky získat bohatý dokladový materiál z živého jazyka pro různě formulované lingvistické otázky na výskyt jevů v jazyce produktivních spolu s jejich lingvistickým hodnocením. Lze získat samozřejmě i představu o zastoupení jednotlivých jevů, tj. je umožněno odlišit jevy centrální (produktivní) a jevy okrajové (periferní, okazionální). Otázky týkající se např. vztahu aktiva a pasíva (popř. druhů pasíva) v češtině, jak často je u pasíva opisného vyjádřen původce děje ap., jaké větné funkce má český infinitiv atd. mohou být zajímavé nejenom pro lingvisty, ale i pro překladatele, učitele ap. PDT se používá jako zdroj dat i při formulaci algoritmických procedur strukturně založené syntaktické analýzy, která následně slouží pro automatické indexování textů, strojový překlad nebo pro gramatický korektor.

5 Závěr

Jak jsme se pokusili ukázat, pomocí matematického aparátu lze modelovat i tak nepravidelný systém, jakým je přirozený jazyk; zároveň lze se slušnou úspěšností tento model konkrétně aplikovat. Kladli jsme zde důraz na statistické metody a demonstrovali jsme jejich použití pro strojové učení. Z hlediska počítačové lingvistiky jsou statistické modely stále ještě nedostatečné.

Pro "ideální" trénování bychom potřebovali taková vstupní data, která obsahují veškeré jevy v takových proporcích, aby umožnily dobré vyvážení potřebných pravděpodobností. Tomu však brání nestatistická povaha jazykových dat: Frekvence slovních tvarů a konstrukcí zdaleka nejsou obrazem jejich skutečného systémového vztahu, a tak může nastat záměna jevů centrálních a okrajových. Je tedy třeba hledat další metody a ověřovat různé přístupy. V oddíle 1 až 4 jsme ilustrovali současný stav výzkumu i jeho výsledky. V budoucnu je třeba hledat další metody a nové cesty (např. vytvářením "hybridních" modelů, srov. odd. 1(e)).

Reference

[1] Alla Bémová a kol. (1997). Anotace na analytické rovině - příručka pro anotátory. *Technická zpráva TR-1997-04*, Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha.

[2] William Byrne, Peter Beyerlein, Juan Manuel Huerta, Sanjeev Khudanpur, Bhaskara Marthi, John Morgan, **Nino Peterek**, Joe Picone, Dimitra Vergyri, Wei Wang (2000). Towards Language Independent Acoustic Modeling. In *Proceedings of the ICASSP2000*, pp. 40-44, Istanbul, Turecko.

[3] Jan Cuřín, Martin Čmejrek (1999). Automatic Translation Lexicon Extraction from Czech-English Parallel Text. In *The Prague Bulletin of Mathematical Linguistics*, vol. 71, pp. 47-57. Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha.

[4] Jan Hajič (1998). Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajičová (Ed.): *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová,* pp. 106-132. Karolinum, Praha.

[5] **Jan Hajič**, Eric Brill, Michael Collins, **Barbora Hladká**, Douglas Jones, Cynthia Kuo, Lance Ramshaw, Oren Schwartz, Christopher Tillmann, **Daniel Zeman** (1998). Core Natural Language Processing Technology Applicable to Multiple Languages. In *Technical Report, NLP WS'98.* Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, USA.

[6] Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall (1998). Syntax v českém národním korpusu. In *Slovo a slovesnost,* vol. 59, pp. 168-177.

[7] Jan Hajič, Barbora Hladká (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset. In *Proceedings of COLING-ACL Conference*, pp. 483-490, Montreal, Kanada.

[8] Jan Hajič, Kiril Ribarov (1997). Rule-Based Dependencies. In *Proceedings of the Workshop on the Empirical Learning of Natural Language Processing Tasks*, pp. 125-136, Praha.

[9] Eva Hajičová (1996). The Past and the Present of Computational Linguistics at Charles Univeristy. *Technická zpráva TR-1996-01*, Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha.

[10] Eva Hajičová, Jarmila Panevová, Petr Sgall (1998). Language Resources Need Annotations To Make Them Really Reusable: The Prague Dependency Treebank. In *Proceedings of the First International Conference on Language Resources & Evaluation,* pp. 713-718. Granada, Španělsko.

[11] Eva Hajičová, Jarmila Panevová, Petr Sgall (1999). Manuál tektogramatického značkování. *Technická zpráva TR-1999-07*, Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha.

[12] Barbora Hladká (2000). *Czech Language Tagging.* Doktorská práce, Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha.

[13] Vladislav Kuboň, Tomáš Holan, Martin Plátek (1997). A Grammar-Checker for Czech. *Technická zpráva TR-1997-02*, Ústav formální a aplikované lingvistiky, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha.

[14] Yaser Al-Onaizan, **Jan Cuřín**, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah A. Smith, David Yarowsky (1999). The Statistical Machine Translation. In *Technical Report, NLP WS'99.* Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, USA.