Korpusy a počítačová lingvistika

(volný překlad dle : ***Tony McEnery & Andrew Wilson:*** **Corpus Linguistics,** Edinburgh Teextbooks in Empirical Linguistics 1996, 1997)

Úvod

Za posledních 10 let došlo v počítačové lingvistice k dramatickým změnám. Od svých počátků se tato disciplína snažila vytvořit plausibilní kognitivní systém na základě introspekce. Nyní se se objevují snahy o vytvoření systému založenému na empirických datech.

V této přednášce nepodáme vyčerpávající úvod do počítačové lingvistiky založené na korpusech. Bylo by k tomu třeba úvodu do teorie informace a statistiky. Pokusíme se o pouhý srozumitelný přehled všech oblastí NPL, pro něž se korpusy staly relevantním východiskem.

O co se tedy budeme snažit?

Pokusím se vás v obecných rysech seznámit s tím, jak a kde se v NPL používá korpusů. Pokládejte tuto přednášku za pouhý úvod do dané problematiky.

Z oblastí NLP, jimiž se dnes budeme zabývat v souvislosti s KL, jmenujme tedy následující: analýza slovních druhů (POS, morfologická analýza), automatizovaná lexikografie, parsing neboli automatická syntaktická analýza a strojový překlad. Budeme se zabývat tím, jak se v těchto jmenovaných oblastech různé techniky využívané specifickými systémy používají korpusy k dosažení některých cílů. Dříve než se ovšem budeme zabývat jednotlivými případy, podívejme se, čím přispěly korpusy NLP jako celku.

Co mohou korpusy nabídnout?

V oblasti počítačové lingvistiky a umělé inteligence se často odkazuje k rozdílu mezi kognitivně plausibilním a kognitivně neplausibilním systémem. Kognitivně plausibilní systém usiluje o vytvoření poznávacího (kognitivního) modelu, pro nějž je relevantní, jakým způsobem člověk řeší nějaký úkol pomocí inteligence, a používá jej jako bázi pro stroj, který má tento problém (úkol) inteligentně řešit. Takový systém často používá úplnou sadu pravidel, které explicitně formulují znalosti, které člověk implicitně používá, ve formě tzv. báze znalostí (knowledge base). Naopak systémy, které rezignují na kognitivní plausibilitu a jsou tedy kognitivně neplausibilním, se prostě snaží vytvořit model inteligentního chování, aniž by se staraly o to, zda systém pracuje stejným způsobem jako člověk (používá inteligenci). Tyto systémy často používají surová kvantitativní data k vytvoření statistického modelu napodobujícího lidské chování.

Korpusy mohou samozřejmě přispět jak k rozvoji kognitivně plausibilních systémů, tak k rozvoji systémů kognitivně neplausibilních. V případě druhých ovšem vzhledem k tomu, že představují zdroje kvantitativních dat, se stávají do značné míry východiskem. Kvantitativné přístupy v oblasti AI (umělé inteligence) byly dosti obtížné dříve než se objevily právě velké korpusy, neboť zatímco člověk je velmi ubohým zdrojem tohoto typu dat, korpus představuje velice spolehlivý a v podstatě nepřekonatelný pramen informací tohoto druhu. V oblasti počítačové lingvistiky mohou tedy korpusy pomoci při vytváření kognitivně plausibilních systémů, pokud se ovšem kognitivní plausibilita obětuje hrubé matematické síle, korpusy představují podmínku sine qua non. Korpusy představují surová data pro veškeré přístupy matematické lingvistiky (modelování na základě matematických, numerických operací).

Není tím nikterak řečeno, že by každý systém, který pracuje s výsledky matematické statistiky, byl automaticky kognitivně plausibilní. Většinou je to pouze otázka stupně. V podstatě existuje jen velmi málo systémů, které by úplně rezignovaly na kognitivně plausibilitu ve prospěch statistických řešení. Obvyklejší je, že se kvantitativní data získaná analýzou korpusů stávají základnou pro systémy NLP založené na tradičních metodách z oblasti AI. Kvantitativní data se většinou používají pro tzv. disambiguace (desambiguaci) – zjednoznačnění víceznačných případů. Podrobněji se k této problematice zaměřím ve výkladu o POS taggování.

V následujícím výkladu se tedy zaměříme na to, jak korpusy přispívají k řešení úloh v oblasti NLP. Ukážeme, kde jsou zdrojem pro různé systémy, které v různém stupni obětují kognitivní plausibilitu a jak přispívají k rozvoji systémů založených na modelech fungujících na bázi tradičních pravidel (kognitivně plausibilní systémy).

Automatická morfologická analýza (POS – analýza)

Značkování korpusu na úrovni slovních druhů (taggování) je nejběžnějším typem značkování a prvním krokem k automatické analýze textu na vyšších rovinách popisu jazyka. Proto je také prioritou pro korpusovou lingvistiku, neboť zásadním způsobem pomáhá při další analýze textu. Stupeň automatizace je procesu značkování na úrovni SD (POS) je zatím dosti vysoký. Existuje celá řada programů pro angličtinu. Situace u nás – dva automatické morfologické analyzátory – taggery, pracuje se na vývoji software automatické desambiguace.

Úkolem taggování na úrovni SD je zařazení slova do odpovídající třídy. Tradiční rozlišování (sb., adj., verb.,..) bylo rozšířeno o další relevantní informace (gramatické významy GK podle SD – rod, číslo, pád, osoba, čas, způsob, vid, typ participia, negace, stupeň, slovní poddruhy atd.). Počet těchto kategorií se liší od systému k systému a je podmíněn metodologickými přístupy, teoretickými východisky autora návrhu značek, potřebami potencionálního uživatele korpusu atd. Opět připomínám dvě ze sedmi maxim G. Leeche. Značkování není neomylnou instancí a značkování by mělo být pokud možno co nejméně závislé na speciální lingvistické teorii. Co se týče metodologie existuje vlastně možnost a) manuálního taggování, a) automatického taggování založeného na 1) statistických metodách (stochastických přístupech) 2) na morfologické analýze prostřednictvím systému pravidel (kognitivně plausibilní model toho, jak analyzuje – určuje SD člověk, třeba žák ve škole).

Vzhledem k tomu, že korpusy jsou rozsáhlé, ruční značkování je téměř nemožné (drahé, časově neefektivní, chyby). Tlak na automatizaci je velký. V obl. NLP se tedy vývoj programů pro automatické značkování POS stává prioritou.

Programy: TAGGIT, CLAWS

Využití korpusů k značkování POS – jak se tagguje

Dříve než se podíváme na to, jak se korpusy mohou využít pro POS tagging, podívejme se na to, jak se vlastně korpusy taggují, jak pracují taggery. Mezi jednotlivými programy existuje řada styčných bodů, takže celkově můžeme zobecňovat.

VSTUP – TEXT v PJ

VE SLOVNÍKU jsou IDENTIFIKOVÁNA SLOVA MORF. ANAL.
a je jim přiřazena informace POS

VÍCEZNAČNÉ PŘÍPADY DESAMBIGUCE

VÝSTUP

V případě slov nenalezených ve slovníku existují programy, které se snaží analyzovat neznámá slova.

Podívejme se nyní, jak fungují jednotlivá stádia.

Slovník

MRF slovník – pokud je slovo identifikováno (nalezeno) ve slovníku, je označeno všemi potencionálními značkami. Pro flektivní jazyky – morf. anal.
Čím větší slovník, tím vyšší úspěšnost (otevřenost slovníku).
Morfologická analýza nenalezených slov.
Automatická MA – pokus odtrhnout potenciální koncovku, ověřit, ke kterému slovnímu druhu by mohl tvar patřit, nabídnou k desambiguaci, použít nějaký automatický desambiguační program.
Desambiguace – řešení hádanky na základě kontextu.

Idiomy – značkování nedělitelných víceslovných jednotek
Celá řada jednotek patřících k témuž slovnímu druhu je dekomponována (předložkové výrazy, spojkové výrazy, citoslovečné výrazy), jindy je záhodno označit více jednotek jako jeden celek (víceslovná pojmenování). Obojí případ se řeší na bázi speciální lexikonu víceslovných jednotek.
Problém – nespojité jednotky

Desambiguace
Desambiguace řízená pravidly. Lingvista na základě znalostí o pravidlech a na základě znalosti postupů, které používá člověk, když provádí odpovídající analýzu, buduje popis (algoritmus), který se může stát bází počítačového programu.
Častější je probabilistický přístup. Programy využívají toho, že některé interpretace víceznačných jevů jsou častější než jiné. Halliday tvrdí, že lidská gramatika je do urč. míry probabilistická. Probabilistické modely vycházejí ze statistik souvýskytu. Slovní okolí (zapojenost do struktury) je určující pro zjednoznačnění významu. Tak např. před substantivem předchází člen nebo adjektivum, slova se řadí podle určitého pořádku do vět nebo frází (pevný slovosled).
Značkování POS – největší prostor pro probabalistické přístupy.
Typologická odlišnost jazyků – velké problémy pro přenos přístupů.
Jak může značkování korpusu přispět KL?
Navržený systém značek – neodpovídá možnostem automatizace – zjednodušení.
Navržený systém značek lze zjemnit, protože analýza ukáže více.
Porovnání obou přístupů – zdá se, že kvantitativně založené modely jsou rychlejší a jednodušší, modely založené na pravidlech se budují pomalu a ačkoliv jsou v jednotlivostech přesnější, celá řada případů nelze formálně popsat – algoritmizovat.
První typ nevyžaduje lidské zásahy, je levnější. Pokud se dopouští chyb, je konzistentní.
Chybuje se ale hodně. Co s tím?

Co s chybami?
Potřeba vyvinout software, který by opravoval chyby automaticky, protože lidská práce je drahá a pomalá (a navíc mnohdy chybná).

Automatické sestavování slovníků
Na korpusové bázi se budují:
1. Počítačové slovníky pro POS taggery
2. Jednojazyčné slovníky
3. Vícejazyčné slovníky
4. terminologické banky

V souvislosti korpus – slovník si můžeme položit otázku, proč se dívat do korpusu, když se můžeme opřít o intuici. Odpověď je nasnadě. Korpus je spolehlivým, rychlým a nezávislým zdrojem informací.

Taggování

Nad korpusy pracují systémy NLP, které umožňují generování slovníků. Nad anotovanými korpusu lze vytvořit slovníky neoznačkovaných slov a z nich doplňovat stávající slovníky, nad nimiž pracují automatické taggery. Tak vlastně korpusy přispívají k vylepšování nástrojů, kterými jsou dále zpracovávány.

Výkladové slovníky

Existence značkovaného korpusu je velkým přínosem pro lexikografa. Kvantitativní přístup k datům je důležitým krokem pro vytváření informačních zdrojů.

MI

1. mezi dvěma slovy patřícími k sobě, tvoří víceslovnou jednotku
2. mezi dvěma slovy, které mají blízký význam

Překladové slovníky

Existence vícejazyčných korpusů a allingement je neocenitelným zdrojem při budování vícejazyčných slovníků. Přiřazení slov v různých kontextech na straně jedné a různé překlady téhož slova na straně druhé činí z tvorby vícejazyčných slovníků úkol neskonale snazší, než tomu bylo doposud. Umožní zařadit patřičné příklady.

Jsou důležité pro aplikace v oblasti strojového překladu.

Terminologické banky

U terminologie je význam 1:1

Paralelní korpusy – neocenitelnou službu.

Korpusy a lexikografie

¨Jak už jsme si řekli korpusy – průlom v obl. lexikografie. Spousta slovníků – na bázi korpusů.

Parsing (automatická syntaktická analýza)

Abychom mohli hodnotit roli korpusů, podívejme se, jak jsou systémy automat. synt. anal. budovány:

1. identifikace slov ve větě
2. přiřazení syntaktické interpretace (funkce)
3. jak se slova pořádají ve fráze
4. správná interpretace vyšších jednotek (frází, klauzí, vět,…)

Hlavním cílem je, jak vidno, identifikace strukturně vyšších jednotek než jsou slovní tvary a nižších než věty. Bohužel představují pokusy o automatizaci poměrně skličující pohled. Automaticky se dá zachytit 30-40%, což je ve srovnání s POS (až 77%) žalostné. Ukazuje se, že jde o mnohem komplexnější roblém a že jeho řešení bude značně složitější.

Podívejme alespoň na různé přístupy, které se uplatňují. Podíváme-li se na možnosti uchopení problému, zdá se že systémy založené na pravidlech mají daleko lepší perspektivy, než probabilisticky orientované analyzátory. Celá řada přístupů ovšem obě metody kombinuje.

Korpusu, obsahující kvantitativní data, jsou především inspirací pro řešení otázek nekvantitativními metodami.

Tradiční gramatiky

Většina systémů parsingu v rámci počítačové lingvistiky je založena na jazykovědných formalismech, pomocí nichž se popis syntaktických vztahů převádí do formy algoritmů. V obl. AI – PROLOG, vhodný formalismus, přítulný uživatelům lingvistům. Omezení není schopen pracovat s masou dat.

Problém při tvorbě pravidel založených na neformálním popise resp. převodu do formalismu je zachování konzistence systému. Dalším omezením je množství pravidel ve formě výčtu.
Korpusy mohou přispět jako trénovací a ověřovací báze.

Radikální statistické gramatiky
Opačný extrém. Kognitivně neplausibilní, vyhýbají se zavádění lingvistické intuice
Snaží se za použití statistických modelů popsat a určit struktury uvnitř jazyka. Kromě anotovaného korpusu nepracují s žádnými meta-znalostmi , které by napodobovaly lidskou intuici nebo pravidla popsaná v tradičních gramatikách. Systém má k dispozici rozsáhlé stromové banky a snaží se z těchto „lesů“ vybrat „strom“, který nejvíce odpovídá předpokládané struktuře analyzované jednotky.
Korpusy představují nutný zdroj statistických údajů.

Hybridní přístupy – kombinace pravidel a statistických metod
Kombinace.
Nejdříve se ručně označkuje část korpusu – trénovací korpus pro statistické přístupy.

Gramatiky nezakládající se na kvantitativních analýzách
Korpus slouží jako zdroj informací o tom, které případy se vyskytují.

Strojový překlad
Strojový překlad a výzkumy zaměřené tímto směrem se vzrůstající měrou zajímají o korpusy, především o paralelní korpusy s allingmentem. Přitažlivost korpusů pro MT se ozřejmí, když se zamyslíme nad AI metodami, o nichž jsme se zmínili. Jedním z rozdílů mezi algoritmikou a inteligenci je použití systému pravidel na str. jedné a sumě znalostí vnějšího světa na straně druhé. MT představuje pokus o napodobení procesu překládání. Jde o aplikaci pravidel pro převod textu z jazyka A do jazyka B prostř. pravidel, která popisují jednotky A a B a vztahy mezi nimi. Překladatel ovšem používá nejen znalostí o systému jazyka, ale i znalostí mimojazykových, znalostí o světě, prostředí, kultuře.... Tedy znalostní bázi vnějazykového kontextu. A právě korpusy představují lákavou nabídku takovou bázi vědomostí jistým způsobem suplovat.
Konkretních aplikací ovšem není mnoho. Komplexnost problematiky zatím způsobuje, že jedn. výsledky nejsou příliš uspokojivé. Jistě všichni znáte překlady typu out of eyes out of mind jako slepý idiot, nebo windows okno, počítačová stanice sun slunce apod.
Myšlenka, že by bylo možné z korpusu čerpat znalosti vzužitelné systémz MT je velmi lákavá. Konkrétní příklad udělal T. McEnery s vzužitím paralelního korpusu anglicko francouzského, který použil jako podporu překladového programu.

Statistický překlad
Statistický překlad představuje velmi vzácnou formu skutečně radikální počítačové lingvistiky založené na kvantitativním přístupu. Docela rezignuje na existující paradigmata MT založená na AI a zcela vylučuje kognitivně plausibilní postup. Tradiční transfer a pravidla převodu, jak je známe z tradičního MT se nahrazují postupem, při kterém se propočítává souvýskyt a pravděpodobnost výskytu založená na datech získaných z korpusu. Statistikové cynicky tvrdí, že hlavní problém nastane, kdzž do týmu přiberou lingvistu. Ten je totiž začne upozorňovat nejen na úskalí, na něž úspěšně naražejí, ale i na ta, na něž dosud nenarazili (náhodou).

EBMT – strojový překlad založený na vzorcích
Nago (1984), Sadler (1989).

EBMT strojový překlad je založen na tom, máme-li paralelní allingmentem propojené korpusy A a B, kde každé větě x v jazyce A je přiřazena věta x v jazyce B. Narazí-li pak strojový překladač z jazyka A do B na větu x, kterou najde ji v korpusu A, pak ji nahradí větou x z korpusu B.
Taková aplikace korpusů představuje velmi silný nástroj a je jistě ohromným lákadlem. Limity takového přístupu jsou ovšem každému, kdo se hloubě+ji zabýval překladatelskými problémy, patrné. Ať by byl použitý paralelní korpus sebevětší, nikdy nebude dost velký, aby pokryl všechny požadavky. Věta není poslední instancí při překladu významu. Je zapojena do širších kontextů, které mohou značně měnit interpretaci téže věty a tím i její potencionální překlad.
Nicméně představuje tento přístup k MT paralelu hybridního přístupu v tom smyslu, že kombinuje statistiku s přístup napodobující lidské chování a je tedz také jistým mezistupněm mezi kognitivně plausibilním a neplausibilním přístupem.

Korpusy nicméně jsou velkým přínosem pro oblast MT. I návrháři překladatelských programů založených na pravidlech mohou velmi zhusta využít celou řadu informací, které mohou získat právě a jen z korpusu, což se taky děje. Existence a rozvoj paralelní allingovaných korpusů je pak nutnou podmínkou pro statistický překlad a EMBT.

Závěr

Co tedy mohou korpusy nabídnout počítačové lingvistice?
Naznačili jsme, že KL a PL spolu velmi úzce souvisejí a že se navzájem doplňují. Korpusy jsou jedinečným zdrojem kvantitativních dat a vědomostí o fungování jazyka.
Nejrozšířenější aplikací jsou programy automatického značkování a automatické desambiguace. Použití v PL roste a prohlubuje se, takže dnes je existence PL bez KL nepředstavitelná.