Akvizícia gramatiky z korpusu pomocou samoorganizácie

Diplomová práca

Miroslav Ľos

Univerzita Komenského v Bratislave
Fakulta Matematiky, Fyziky a Informatiky
Katedra Aplikovanej Informatiky

Študijný odbor: Informatika

Vedúci diplomovej práce: doc. Ing. Igor Farkaš, PhD.

Bratislava, Máj 2007

Poďakovanie

Týmto sa chcem poďakovať v prvom rade svojmu diplomovému vedúcemu Igorovi Farkašovi za výber témy práce a poskytnutie odbornej literatúry, ale hlavne za jeho otázky, ktorými ma viedol k presnejšiemu skúmaniu a pochopeniu problematiky.
Ďalej sa chcem poďakovať menovite Radovanovi Garabíkovi, ale aj celému Slovenskému Národnému Korpusu JÚĽŠ SAV za vytvorenie a sprístupnenie korpusu textov súčasného slovenského jazyka a poskytnutie potrebného hardvérového vybavenia.
Nakoniec sa chcem poďakovať svojím kolegom v ročníku, keď svižnosť ich postupu ma najväčšmi poháňala pri vypracovávaní tejto práce.

Čestne prehlasujem, že som túto diplomovú prácu vypracoval samostatne s použitím uvedenej literatúry.

V Bratislave, 3. mája 2007 .....
Miroslav Ľos, v.r.

Abstrakt

V práci je popísaný model získavania gramatiky viet z čistého neanotovaného textového korpusu vyhľadávaním signifikantných kontextovo závislých vzorov štruktúry viet korpusu prostredníctvom redukcie redundancie systematickou unifikáciou čiastočne zhodných ciest v grafovej reprezentácii viet korpusu.
Tento model je následne rozšírený o metódu vyhľadávania význačných morfém jazyka bez učiteľa v korpuse rozdelenom na jednotlivé znaky, pričom sa vyhýba nutnosti tvoriť vzory pod úrovňou slov bez sémantickej plauzibility a syntaktickej opodstatnenosti.
Na pokusoch s náhodne generovanými umelými gramatikami a korpuse textov slovenského jazyka sa preukážu prezentované schopnosti tohto modelu.
Kľúčové slová: Akvizícia jazyka, štatistické spracovanie prirodzeného jazyka, učenie bez učiteľa, konštrukčná gramatika

Abstract

In the present work a grammar acquisition model is described that searches for significant and context–specific patterns in the structure of the sentences found in a raw, unannotated corpus, utilizing redundancy reduction by systematic unification of partially aligned paths in a graph–formed representation of the corpus.
This model is then enhanced with an unsupervised search method for significant morphemes of the language in a corpus split–up to sole characters while avoiding the need to produce sub–word–level patterns without any semantic plausibility and syntactic justifiability.
With experiments on random–generated artificial grammars and a corpus of texts in Slovak language the advertised capabilities of this model shall be shown.
Keywords: Language acquisition, statistical natural language processing, unsupervised learning, construction grammar

Obsah

1 Úvod

Prirodzený jazyk je jedným z najzaujímavejších objektov výskumu v umelej inteligencii a príbuzných vedeckých disciplínach. Primäť umelé systémy skutočne porozumieť výrokom vyjadreným a vyjadriteľným v prirodzenom jazyku sa považuje za významné, ak nie kľúčové kritérium na uznanie takéhoto systému inteligentným. Jednotlivé zložky či podproblémy tejto úlohy sa niekedy považujú za natoľko fundamentálne späté s inteligenciou, až vznikla metafora o AI–úplnosti porozumenia prirodzenému jazyku, čiže myšlienke, že rozriešenie tejto úlohy bude obsahovať, alebo mať za následok aj riešenia mnohých iných (aj keď možno nie všetkých) problémov silnej UI.

Prirodzeným zdrojom modelov či aspoň myšlienok riešiacich (pod–) problémy porozumenia prirodzeného jazyka sú ľudia ako jeho (jediní súčasní) používatelia. Ľudia majú skoro bez výnimky schopnosť naučiť sa jazyk svojej komunity v povážlivo krátkom čase, pričom sa ich prejavované jazykové prostriedky navzájom líšia len v nepatrnej miere. S pomocou jazyka dokážu teoreticky formulovať, uchovávať a komunikovať všetky svoje myšlienky na ľubovoľnej úrovni. Komunikácia takýmto spôsobom s umelými systémami by mala nespochybniteľný efekt kvalitatívneho pozdvihnutia práce s nimi. Konečnou úlohou teda je vytvoriť systém schopný vyextrahovať význam uložený vo vetách nejakého (ľubovoľného) jazyka a obrátene utvoriť ku komunikovaným informáciám vety jazyka spôsobom, akým by ich utvoril človek. Táto úloha teda spadá do skupiny úloh UI snažiacich sa napodobniť ľudské výkony. Prakticky všetky úlohy z tejto kategórie sa veda prirodzene snaží riešiť tým, že najprv preskúma spôsob, akým ich dokáže riešiť človek.

Kým jedna časť lingvistov pokladá úlohu naučenia sa jazyka za podmienok, za akých zjavne prebieha za príliš náročnú a predpokladá, že všetky prirodzené jazyky majú spoločný jednotný základ, ktorý je spracovateľný na to uspôsobeným modulom v mozgu, alternatívna komunita nepovažuje argumenty týchto tzv. nativistov za dostatočné a snaží sa hľadať všeobecné učiace postupy, ktoré nie sú špeciálne zamerané na jazyk, avšak dokážu získať jeho použiteľnú reprezentáciu. Takéto metódy sa týmto vlastne pokúšajú na dostatočne zložitom probléme nájsť aj jednotný mechanizmus podobný tomu, ktorý sa zasluhuje o schopnosti človeka pripisované inteligencii.

Metódy, ktoré využíva táto druhá skupina sú založené na pozorovaní pravidelností v množine prístupných pozorovaní, teda pozorovaním štatistických závislostí medzi jednotlivými slovami a vetami v korpuse viet daného jazyka. Samostatné naučenie sa gramatiky jazyka umelým systémom by v porovnaní s „ručným“ prístupom jej naprogramovania znamenal neporovnateľný kvalitatívny prínos pre umelú inteligenciu, pričom je z jej pohľadu aj zaujímavejší, čo je dôvodom, pre ktorý som si vybral takýto model za tému svojej diplomovej práce.

V druhej kapitole tejto práce popíšem problematiku teórií učenia sa jazyka, prostriedky a metódy, ktoré využíva štatistický blok spracovania prirodzeného jazyka (anglicky [Statistical] Natural Language Processing – NLP) a základné poznatky zo štatistiky a teórie informácie, na ktorých sa zakladajú spomenuté metódy. Tretia kapitola sa zaoberá modelom na automatické získavanie štruktúry z korpusu, algoritmom, na ktorom je založená táto práca. Štvrtá kapitola opisuje nedostatky a praktické výhrady voči Adiosu, ktoré čiastočne vznikajú pri jeho aplikácii na slovenský jazyk a návrhy na ich riešenie. Piata kapitola opisuje experimenty vykonané s navrhnutým modelom a analyzuje dosiahnuté výsledky, ktoré sú zhrnuté v závere práce.

2 Konceptuálne a teoretické východiská

2.1 Gramatika prirodzeného jazyka a jej učenie

Význam myšlienok človeka sa do viet prirodzeného jazyka ukladá usporiadavaním (v každom momente konečnej) množiny symbolov do konečných postupností. Toto usporiadavanie sa vyznačuje vysokým stupňom systematického štruktúrovania v opakujúcich sa vzoroch (pattern, nie (len) v zmysle deklinačného vzoru). Tieto konzistentne používané vzory v sebe uchovávajú komponenty zamýšľaného významu. Spôsob, akým na seba vzory môžu nadväzovať môže nadobúdať podobu (produkčných) pravidiel. Pravidlá, ktorými sa produkuje a zachytáva význam vo vetách, tvoria gramatiku jazyka. Ovládnutie schopnosti komunikovať v jazyku teda znamená ovládnutie, získanie pravidiel jeho gramatiky. Deti (aj dospelí) sa gramatiku daného konkrétneho jazyka (aspoň v jej konečnej podobe) musia učiť, keďže dopredu nemajú možnosť poznať tento jazyk.

Nativizmus a generatívna gramatika

Niektoré poznatky z teórie učenia (najviac spomínané sú asi z (Gold, 1967)) naznačovali, že niektoré v gramatikách prirodzených jazykov pozorovateľné javy sa nedajú naučiť zo vstupu (pozitívnych príkladov teda vypočutých viet jazyka), ktorý deťom celkom bežne stačí na perfektné zvládnutie týchto javov, ako aj celého jazyka. Z tohto skupina vedcov na čele s Noamom Chomskym označovaná ako nativisti usúdila, že ľudia musia mať vrodené jazykovo–špecifické schopnosti zakódované v nejakej časti mozgu. Na tejto myšlienke bola potom postavená teória Univerzálnej Gramatiky, teda jadra gramatiky spoločnej všetkým ľuďom. V tejto gramatike resp. v tzv. hlbokých štruktúrach je reprezentovaný celý význam komunikovanej myšlienky. Všetky rôznorodé gramatiky prirodzených jazykov potom vznikajú pridaním (či vlastne aktiváciou a nastavením parametrov) transformačných pravidiel pre prepis hlbokej štruktúry na plytkú, teda prezentovanú štruktúru jazyka. Učenie sa jazyka sa takto redukuje na hľadanie vhodných parametrov pre transformačné pravidlá a získanie (zapamätanie si) lexikónu slov (najmenšiemu druhu vzorov v zmysle predchádzajúceho odseku) a ich priradených významov (Sag, Wasow 1998).

S týmto nativistickým pohľadom na gramatiku je prirodzene spätý aj používaný tvar pravidiel gramatiky: Vety sa členia do (usporiadaných) fráz a podfráz (ako substantívna, slovesná, predložková), ktoré zase určujú pozície a kategórie použiteľným slovným druhom, za ktoré možno dosadiť vhodné slová z lexikónu. K takýmto pravidlám sa pridali pravidlá, ktoré dokázali transformovať základnú vetu do iných podôb, napr. utvárajúc trpný rod, otázku, imperatív či substantíva zo slovies. Nativisti tak založili rodinu generatívnych gramatík. Tieto následne prešli viacstupňovým vývojom, napríklad postupne sa zbavujúc transformačných pravidiel v prospech pravidiel založených na ohraničeniach neutrálnejších oproti pôvodným postojom nativistov.

Asi najpoužívanejšou triedou súčasných generatívnych gramatík je Head–Driven Phrase Structure Grammar (HPSG). Triedy fráz, z ktorých pozostávajú vety tvoria typovú hierarchiu obohatenú o hierarchiu vlastností (features; ide o syntaktické aj sémantické kategórie fráz) a možné hodnoty týchto vlastností (ide teda o ručne konštruovanú komplikovanú rámcovú reprezentáciu poznatkov o gramatike jazykov). Frázy vo vetách vytvárajú znamenia (signs), čo sú dvojice postupnosti slov tvoriacich frázu a čiastočne ohodnoteným rámcom frázy. Jednotlivé prvky frázy majú priradené vlastné rámce. Dôležitým prvkom systému sú hlavy zložených fráz, čo sú ich najdôležitejšie prvky, zvyčajne zdieľajúce hodnoty dôležitých vlastností (napr. hlavou podmetovej frázy v nominatíve singuláru je podstatné meno rovnakých vlastností). Pravidlá gramatiky sú obohatené ohraničeniami na typy a hodnoty ich vlastností, tiež ich môžu do rámcov dopĺňať. Slová z lexikónu majú určené typy a niektoré hodnoty, čím sa definuje ich použiteľnosť vo frázach. Gramatickosť vety sa deklaruje existenciou sústavy znamení konzistentnej s ďalšími princípmi a ohraničeniami popisujúcich (syntaktickú aj sémantickú) úlohu každého slova vety (Sag, Wasow 1998).

Iným pohľadom na (generatívnu) gramatiku jazyka viac kompatibilná so štatistickými náhľadmi je rodina konštrukčných gramatík. V nich sa neodlišujú syntaktické pravidlá a lexikón gramatiky, ale celá gramatika sa berie ako „konštruktikon“ – zbierka konštrukcií, teda pevných syntaktických jednotiek rôznej veľkosti, ktoré predstavujú stavebné kamene na tvorbu viet jazyka. Každá konštrukcia má priradený svoj vlastný význam (je to vlastne pár formy a obsahu), ktorý tak nie je len funkciou významov svojich podkonštrukcií ako inde. Konštrukcie môžu mať formu od morfém a slov, cez idiómy až po abstraktné syntaktické konštrukcie (z generatívneho pohľadu). Tieto sa berú rovnocenne a

utvárajú akési syntakticko–lexikologické kontinuum. Všetky jazykové znalosti človeka sú obsiahnuté v sémantickej sieti jemu známych konštrukcií (Goldberg 2003).

Lingvistické korpusy

Každá vedecká teória sa v konečnom dôsledku potvrdzuje alebo vyvracia na empirických dátach. Za účelom potvrdzovania alebo revidovania názorov jednotlivých lingvistov sa začali takéto dáta zbierať. Empirickými dátami v lingvistike sú skutočne používané (písané či vyslovené) vety, rozhovory či výpovede v skúmanom jazyku. Tieto dáta sa vhodne zoskupujú do korpusov textov.

Textové (alebo hovorené) korpusy podľa zamerania v rámci jazyka zoskupujú texty určitej kategórie, alebo sa naopak snažia zhromaždiť reprezentatívny prierez celým jazykom v danom časovom období. Medzi často používané korpusy prvého druhu patrí korpus CHILDES zachytávajúci rozhovory medzi rodičmi a deťmi. Na tomto korpuse sa zvyčajne robia experimenty s modelmi učenia sa jazyka malými deťmi. Korpusy druhej kategórie sú rôzne národné korpusy, ako napríklad aj Slovenský Národný Korpus.

Súčasťou práce na korpuse zvyčajne býva aj segmentácia textov na tokeny a vety a priradenie rôznych druhov anotácií textom alebo ich častiam, od štýlovo–žánrových cez morfologické, syntaktické až po sémantické či pragmatické (tieto sú skôr vysnívaným cieľom tvorcov korpusov). Azda najznámejším príkladom korpusu so syntaktickou anotáciou je produkt Pennsylvánskej univerzity – Penn Treebank. Tieto anotácie sa ponúkajú ako zlatý štandard pre porovnanie s navrhovanými modelmi (akvizície) gramatiky. Problémom však je, že tieto anotácie odzrkadľujú teoretické názory alebo skúsenosť autorov korpusu s gramatikou, pričom jednotlivé názory sa rôznia už aj pri takých jednoduchých anotáciách ako je uzátvorkovanie (bracketing; zoskupovanie, napríklad prívlastkov k podstatným menám) niektorých konštrukcií líšia v zásadných bodoch (Manning, Schütze 1999).

Štatistický pohľad na jazyk

Vznik korpusov podnietil znovuoživenie štatistických a pravdepodobnostných prístupov k jazyku. Tieto boli totiž s nástupom nativistických teórií potlačené do ústrania. Pred týmto obdobím boli empirici zastupovaní britským lingvistom J. R. Firthom, so známym (obľúbeným anekdotickým a opisným) heslom: „You shall know a word

by the company it keeps.“ (Firth 1957, str. 10) Pre tieto prístupy je príznačná snaha vysvetliť produkciu viet jazyka ako stochastický proces, s určitými pravdepodobnosťami výskytu a poradia slov. Jazyk v takomto ponímaní nie je tak kategoricky odlíšená množina gramatických viet ako v prípade generativistov, ale každá veta môže byť vyslovená s určitou pravdepodobnosťou (mizivou u „negramatických“ viet). Vysoká pravidelnosť vzorov v jazyku sa chápe ako následok výberu podľa sociálne–evolučných tlakov na efektívnu prenositeľnosť a reprezentovateľnosť významov. Oproti nativistickým názorom o prakticky konečnom (obohacuje sa iba lexikón) počte pravidiel každej gramatiky jazyka takéto modely konceptuálne umožňujú dynamické zmeny a obohacovanie či ochudobňovanie gramatiky v čase. Význam slov a fráz sa chápe byť určený svojím zvyčajným (častým, pravdepodobným) použitím. Takéto chápanie jazyka umožnilo odpútať sa od námietok nativistov a zaoberať sa akvizíciou gramatiky jazyka z (pozitívnych) dát obsiahnutých v korpuse. Štatistické prístupy k jazyku sa v teórii opierajú o konekcionistické kognitívne a neurovedné teórie, ktoré poukazujú na schopnosť mozgu „všímať si“ a pracovať na základe štatistických vlastností vnemov. Preto sa štatistické smery akvizície jazyka snažia podporiť tieto teórie výsledkami s modelmi automatického učenia sa gramatiky deťmi bez učiteľa pomocou všeobecne (nielen jazykovo) použiteľných algoritmov učenia.

2.2 Štatistická akvizícia gramatiky a jej metódy

Štatistické modely jazyka odhadujú distribúcie slov a fráz pomocou rôznych štatistík a frekvencií výskytov jazykových javov v korpuse. Najpoužívanejšími sú relatívne frekvencie kookurencií či kolokácií skupín slov vo vzájomnej blízkosti alebo následnosti (podľa dĺžky sekvencie rozlišujeme štatistiky bigramové, trigramové, štvorgramové, atď.). Pomocou nich sa tvoria rôzne kritériá na významnosť (pravdepodobnosť, prípustnosť) kandidátskych vzorov, ktoré sa podľa nich zahrnú do vytváranej gramatiky. Tieto bývajú buď probabilistické, ako rôzne použitia EM (expectation maximisation) algoritmu, ktoré sa snažia nájsť gramatiku z nejakej vopred určenej rodiny, ktorej použitím sa vety korpusu stanú pravdepodobné v maximálnej možnej miere, alebo informačno–teoretické (Minimum description length), pri ktorých sa dáva prednosť vzorom, ktorých použitie minimalizuje nároky na vyjadrenie (a maximalizuje prírastok informácie) najpoužívanejších vetných konštrukcií. Základným modelom používaným na jednoduché získanie gramatiky je PCFG, ktorá klasickým pravidlám bezkontextovej

gramatiky priraďuje pravdepodobnosť ich použitia. Týmto spôsobom sa jednoducho priradí každej generovateľnej vete jej pravdepodobnosť (a tým jej očakávaná praktická použiteľnosť) (Manning, Schütze 1999).

Niektoré modely štatistického učenia sa gramatík

Niektoré zo základných myšlienok konštrukcionisticky ladených modelov gramatík a ich učenia obsahujú Memory Based Learning modely ako FAMBL (Roberts, Atwell 2002). Tento model využíva slabú abstrakciu (klastrovaním metódou k–nearest neighbor) na získanie rodiny podobných výrazov, ktoré sa zlúčia do jedného rodinného vzoru. Tento zachytáva spoločný tvar rodiny, pričom sa pôvodné výrazy reprezentujú ako alternatívy v podvýrazoch tohto vzoru. Schematicky to možno vyjadriť ako nahradenie rodiny výrazov ABCD, ABED, AFCD rodinným výrazom A(B,F)(C,E)D. Ďalšia (podobná) myšlienka je využitá aj v Alignment Based Learning. Cez pravdepodobnosti výskytu slov v konštituentoch sa hľadajú koherentné čiastočne zhodné (partially aligned) konštituenty, ktorých nezhodné časti sa považujú sa toho istého typu a vzájomne zameniteľné, z čoho sa odvodia príslušné pravidlá gramatiky (Roberts, Atwell 2002).

Wolff vo svojej práci (1988) predstavuje teóriu učenia sa gramatiky ako znižovania redundancie kompresiou kognitívnych štruktúr. Podľa nej sa človek učí gramatiku jazyka úpravou pravidiel tak, aby minimalizoval nároky na pamäťový priestor pri zachovaní vyjadrovacej účinnosti gramatiky. Implementáciou niektorých konceptov tejto teórie je v práci predstavený program SNPR V ňom sa všetky jazykové znalosti ukladajú do štruktúry pravidiel. Tieto pravidlá sú troch druhov: minimálne (M, produkujú písmená), syntagmatické (SYN, opisujú sekvencie pravidiel; každé pravidlo expanduje jedinečný neterminál) a paradigmatické (PAR, opisujú alternatívne použiteľné pravidlá). Z frekventovaných párov pravidiel v korpuse sa tvoria SYN pravidlá, ktoré (rekurzívnym užitím) postupne utvárajú frekventované slabiky, slová, slovné spojenia, frázy a vety. Pravidlá so spoločnými kontextami (pravidlami pred a za nimi) sa zoskupujú do komplexných SYN pravidiel, ktoré obsahujú PAR pravidlá na zachytenie rozdielov v rámci skupiny. Takto utvorené PAR pravidlá sa podieľajú na zovšeobecneniach: v pravidlách, ktoré obsahujú niektorú z alternatív PAR pravidla, sa táto nahradí samotným PAR pravidlom. Dáva sa tak možnosť vzniku nových viet výberom inej alternatívy PAR pravidla. Možnosť prílišného zovšeobecnenia sa eliminuje prestavovaním PAR

pravidiel: ak pre niektoré z alternatív PAR pravidla nie je dostatočná podpora v korpuse, nahradí sa tento výskyt pravidla novým PAR pravidlom, ktorý obsahuje len podporované alternatívy. Dobré generalizácie sa týmto nestrácajú, pretože tie majú podporu v korpuse dostatočnú (prakticky sú tým definované). Wolff pokračuje v práci (Wolff 1988) empirickým zdôvodňovaním práce algoritmu a teórie v pozadí (segmentácia, učenie sa slov, skladanie stále zložitejších konštrukcií na základe hľadania zhody v histórii počutých viet, zovšeobecnenie a eliminácia prílišného zovšeobecnenia) z pozorovaní reálneho priebehu učenia u detí. Naoplátku pomocou tejto teórie možno elegantne vysvetliť niektoré ďalšie javy v učení. Poskytovaná teória nie je zameraná výlučne na jazyk – dá sa využiť aj na iné druhy učenia, čo sa chápe ako podpora nie–nativistických pohľadov na akvizíciu jazyka.

Modelom akvizície jazyka využitým a rozvinutým v tejto diplomovej práci je algoritmus ADIOS (popísaný v článkoch Solan et al. 2003a, Solan et al. 2003b, Solan et al. 2004, Edelman et al. 2003, Edelman et al. 2004a, Edelman et al. 2004b). Korpus sa v ňom uchováva v orientovanom grafe s uzlami tvorenými pôvodnými tokenmi a štruktúrovanými vzormi s asociovanými triedami ekvivalencie (porovnateľné s M, SYN a PAR pravidlami z (Wolff 1988)). Vety korpusu sú reprezentované ako orientované cesty v tomto grafe. Algoritmus prebieha na konceptuálnej úrovni obdobne ako u Wolffa, avšak signifikantnosť (a prípustnosť zaradenia) vzorov sa meria pomocou informačných mier na k–gramových štatistikách (určených počtom hrán grafu v danom smere) a prílišnému zovšeobecneniu sa predchádza skeptickým používaním tried ekvivalencie. Ďalším prínosom modelu je možnosť testov gramatickosti viet na základe miery zhody s generovanou štruktúrou. Bližší opis modelu je predmetom kapitoly 3 tejto práce.

2.3 Štatistické nástroje spracovania jazyka

Frekventovaným nástrojom štatistického spracovania prirodzeného jazyka je testovanie hypotéz. Ide o metódu na overovanie hypotéz o pravdepodobnostných distribúciách skúmaných javov pomocou štatistických testov na nameraných vzorkách dát. Dôležitou kategóriou takýchto testov je test pomerom vierohodností (likelihood ratio test). Pri tejto metóde sa formulujú dve hypotézy o skúmanom jave: tzv. nulová hypotéza H a alternatívna hypotéza A. Nulová hypotéza vysvetľuje pozorované javy spôsobom, ktorý chceme vyvrátiť (napr. že pozorované dáta vznikli čistou náhodou). Alternatívna hypotéza v silnejšej verzii

vysvetľuje daný jav ako komplement nulovej hypotézy, teda predpokladá jej neplatnosť, alebo v slabšej verzii nejakým alternatívnym spôsobom, s ktorým chceme nulovú hypotézu porovnať. Na porovnanie týchto dvoch hypotéz sa vypočíta ich vierohodnosť pri pozorovaných dátach, čo je podmienená pravdepodobnosť, že je daná hypotéza pravdivá, ak nastal pozorovaný jav.

Hodnota pomeru vierohodností nulovej a alternatívnej hypotézy je štatistika na pozorovaných dátach, ktorej hodnota rastie so zväčšujúcou sa pravdivosťou či pravdepodobnosťou nulovej hypotézy. Ak je hodnota tejto štatistiky menšia ako nejaká hraničná hodnota α, môžeme nulovú hypotézu zamietnuť a prípadne prijať či dať prednosť alternatívnej hypotéze, keďže predmetom testu je určiť, ktorá hypotéza lepšie vysvetľuje pozorované javy. Hraničná hodnota α je funkciou signifikantnosti teda pravdepodobnosti chyby prvého druhu tohto testu, takže zvyšovaním α znižujeme pravdepodobnosť, že odmietneme nulovú hypotézu, hoci táto je lepšia ako alternatívna.

Variáciou testu pomeru vierohodností je použitie Kullback–Leiblerovej divergencie podľa vzorca:

$$KLD((A|X)\,\|\,(H|X)) = E\left(\log\frac{P(A|X)}{P(H|X)}\right) = \sum_{x\in X} L_A(x).\log\frac{L_A(x)}{L_H(x)}$$

kde X je množina pozorovaní a L_A a L_H sú vierohodnosti alternatívnej resp. nulovej hypotézy pre dané pozorovanie. Kullback–Leiblerova divergencia je mierou z teórie informácie hovoriacou o očakávanej neefektívnosti optimálneho kódovania informácie (hodnôt) predpokladajúc, že tieto hodnoty sú distribuované podľa nulovej hypotézy, ak je alternatívna hypotéza pravdivá. KLD je pri testovaní hypotéz interpretovateľná ako miera toho, o koľko lepšie sa na popis javu hodí alternatívna hypotéza oproti nulovej a pri štatistickom spracovaní vzorov tiež, ako efektívnejšie popisuje A jav oproti H, čo je podstatou MDL metód NLP.

Keďže pravdepodobnosti výskytu jazykových javov, na ktoré sa redukujú funkcie vierohodnosti nie sú známe, je treba ich odhadnúť z nazbieraných empirických dát. Na tento odhad zvyčajne slúžia odhady opierajúce sa o maximálnu vierohodnosť, teda pravdepodobnosť javu sa odhadne tak, aby bola pravdepodobnosť empirických dát pri danej pravdepodobnosti maximálna. Táto pravdepodobnosť je určená relatívnymi frekvenciami (pomerom výskytu daného výsledku oproti súčtu výskytov všetkých možných elementárnych výsledkov)

jednotlivých elementárnych výsledkov, keďže je maximálne pravdepodobné, že nejaký výsledok dosiahne danú relatívnu frekvenciu práve vtedy, keď je táto relatívna frekvencia rovná jeho pravdepodobnosti. Ak túto myšlienku rozšírime aj na hľadanie modelov jazykových javov v NLP, dostávame sa k EM metódam NLP.

Odhady pravdepodobností z empirických dát

V prípade korpusu viet nejakého jazyka sa teda pravdepodobnosť výskytu určitého slova v jazyku reprezentovanom týmto korpusom odhadne ako pomer počtu jeho výskytov a počtu všetkých slov v korpuse. Odhad podmienenej pravdepodobnosti nejakého slova, za nejakej podmienky (napr. že nasleduje po nejakom inom slove) odhadne ako pomer počtu, kedy sa dané slovo nachádza vo vete spĺňajúcej podmienku a celkovému počtu takýchto viet. Odhady súčasných výskytov viacerých javov sa môžu získať pomocou reťazového pravidla pre podmienenú pravdepodobnosť:
$P(A_1, \ldots, A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot \ldots \cdot P(A_n \mid A_1, \ldots, A_{n-1})$.

Veľkým problémom takýchto odhadov býva zrejmý fakt, že korpus jazyka neobsahuje všetky možné slová a vety daného jazyka, je len jeho viac či menej reprezentatívnou vzorkou. Keďže je možných slov veľmi veľa, vyskytuje sa väčšina slov aj vo veľkom korpuse len niekoľko málo krát, čiže každé slovo má mizivú pravdepodobnosť. Pravdepodobnosť sekvencií slov je o to väčším problémom. Maximum–Likelihood odhady tiež priradia slovám, ktoré sa v korpuse nenachádzajú nulovú pravdepodobnosť a slovám v korpuse teda väčšiu pravdepodobnosť, ako v skutočnom jazyku majú. Preto sa zvyknú tieto odhady upravovať tak, že sa aj neprítomným slovám priradí nejaká minimálna pravdepodobnosť, kým slovám prítomným v korpuse sa z pravdepodobnosti proporcionálne uberie. Tieto techniky sa zvyknú nazývať vyhladzovanie. Jednoduchým príkladom takejto úpravy je použitie Laplaceovho odhadu, ktorý predpokladá uniformnú distribúciu možných slov a skutočne sa vyskytujúcim slovám ju zvýši podľa ich počtu:

$$P_L(w) = \frac{C(w)+1}{N+B}$$

kde C(w) je počet výskytov slova (či iného „javu“), N je počet slov v korpuse a B je počet všetkých možných slov, čo je aj problém tejto metódy: B je potrebné vedieť (alebo zvoliť) a ak je jeho hodnota omnoho väčšia ako N, tak sa nevyskytujúcim sa slovám a hlavne n–gramom priradí väčšia časť (väčšina) distribučnej masy, ako je

z empirických dát v korpuse opodstatnené. Preto sa vyvinuli metódy viac sa opierajúce o štatistiky korpusu. Vyzdvihovanou je napríklad Good–Turingov odhad, ktorý v sebe skrýva predpoklad o binomickom rozdelení početností n–gramov (napriek tomu, že je takýto predpoklad v podstate tiež nesprávny metóda dosahuje akceptovateľné výsledky) a pri odhadoch tak využíva informáciu o početnostiach početností (r a N_r) n–gramov v korpuse a vybranú, tzv. vyhladzovaciu funkciu S (napr. hyperbolickú funkciu $a \cdot r^{b < -1}$, kde parametre sa určujú podľa N_r): odhad javov s početnosťou r menšou ako nejaká konštanta sa upraví pomerom s početnosťou javov o 1 početnejších (napr. pri tzv. Jednoduchom G–T odhade a $r = 0$ je $P_{GT} \approx N_1 / N_0 N$) a potom sa renormalizuje (aby bola suma P jedna) (podrobnejšie v Manning, Schütze 1999).

3 Model akvizície gramatiky ADIOS

3.1 Princípy a ciele modelu

Základným princípom, ktorý sa model Adios (teda Automatic Distillation of Structure) snaží sledovať je chápanie osvojovania jazyka ako hľadanie štruktúry v dátach. Preto sa odkláňa od tradičných generatívnych systémov gramatík v prospech štatistického spracovania jazyka. Za hlavnú nevýhodu umelo konštruovaných generatívnych gramatík prirodzených jazykov sa považuje (hoci prirodzená) snaha lingvistov skonštruovať gramatiku, ktorá odráža ich skúsenosť so systematizovaním poznatkov, a preto sú orientované na zjednotenú hierarchizáciu konštrukcií (istý pohľad zhora nadol), takže tvorba viet predchádza od abstraktných a globálne použiteľných konštrukcií k ich jednotlivým konkretizáciám. Takéto gramatiky ale nemusia byť adekvátne vzhľadom na skutočnosť, že prirodzené jazyky sa vyvinuli (a vyvíjajú) spôsobom analogickým s biologickou evolúciou a neboli zámerne stvorené na svoj účel (prenosu informácie) podľa nejakého plánu. Na druhej strane sú poznatky o jazykových javoch zozbierané lingvistami v jazykoch reálne obsiahnuté, a preto je vhodné a aj potrebné ich vysvetliť a vybudované formalizmy sú vskutku vhodné na reprezentáciu ľubovoľnej informácie.

Aby bol cieľ vytvoriť systém, ktorý hľadá štruktúru v dátach dôsledný, musí sa zbaviť predpokladov zo strany autorov či lingvistov o danom jazyku. Skutočne, hľadanie gramatiky jazyka by nemalo byť spojené s jej predurčovaním. Preto nemožno takémuto systému poskytovať k čistému textovému korpusu (či už písanému alebo reprezentujúcemu prepísanú reč) dodatočné informácie. Príkladom takejto nevhodnej informácie je anotovanie jednotlivých slov ich klasifikáciou (napríklad do slovných druhov, ale aj podrobnejšie), hoci práca na takto anotovaných dátach (či priamo dátach nahradených ich anotáciami) môže prekonať problémy s nízkymi relatívnymi frekvenciami slov a ich spojení v korpuse. Takáto klasifikácia obmedzuje možnosť získať jemnejšiu klasifikáciu na základe lokálnych, kontextovo závislých pravidiel.

Hľadanie štruktúry sa neobmedzuje len na vyhľadávanie a vyberanie takých pravidiel, ktoré sú najlepšou reprezentáciou daného korpusu viet či už z hľadiska pravdepodobnosti alebo popisnej zložitosti, ako to možno vidieť u modelov využívajúcich maximalizáciu očakávaní, resp. minimalizácie dĺžky popisu. Dobrá gramatika získaná z daných dát totiž

nie je tá, pre ktorú je pravdepodobnosť daných dát najväčšia (a z tohto hľadiska je skutočne najpravdepodobnejšia), ani tá, ktorá je efektívna a pritom najjednoduchšia, teda dokáže generovať každú vetu použitím minimálneho počtu pravidiel, ale tá, ktorá dokáže nájsť pravidlá schopné zovšeobecnenia daného fragmentu na celý jazyk. Takéto pravidlá majú nádej reprezentovať javy jazyka spôsobom prelínajúcim sa s významom viet jazyka obsiahnutým priamo v jeho syntaxi.

Reprezentácia gramatiky pomocou vzorov

Každá gramatika je v konečnom dôsledku obmedzená možnosťami, ktoré jej poskytuje jej zvolená reprezentácia. Preto je vhodné voliť takú reprezentáciu, ktorá je kompatibilná s výsledkami psychologických štúdií na ľuďoch. Tieto zdôrazňujú centrálny význam vzorov, označovaných ako konštrukcie, pri reprezentácii aj učení sa jazyka deťmi z príkladov na základe štatistických náznakov. Vhodnými sa teda ukazujú štruktúry zastrešené konštrukčnými gramatikami, ktoré sú jednoducho inventárom lingvistických jednotiek z celého spektra zložitosti. V tejto reprezentácii sú globálne pravidlá výnimkou, forma gramatiky sa skôr skladá z veľkého množstva pevných konštrukcií, z ktorých sa „vylovia“ akékoľvek zovšeobecnenia, aké sa nájdu. Takéto pravidlá sú v porovnaní so všeobecnými pravidlami klasických gramatík použiteľné len v konkrétnych (obmedzených) prípadoch, na druhej strane je ich použitie bezpečnejšie z hľadiska možnosti prílišného zovšeobecnenia.

Reprezentácia gramatiky musí umožniť existenciu konkrétnych, ako aj celkom abstraktných konštrukcií, ktoré sú nezavrhnuteľnou možnosťou reprezentácie významných čŕt jazyka. Tieto abstraktné pravidlá by avšak nemali byť oddelené od reprezentácie konkrétnych konštukcií či samotných slov, napríklad vo forme meta–jazyka (ako je tomu pri HPSG), ale byť hladkým rozšírením slov a slovných spojení.

Takýmto hladkým rozšírením môžu byť vzory. Najkonkrétnejším typom vzorov sú znaky, slabiky, kmene a morfémy či celé slová. Prvým stupňom rozšírenia takýchto vzorov je ich usporadúvanie do postupností, čím vznikajú konkrétne konštrukcie ako kompozitá (zložené slová) a idiómy (ustálené slovné spojenia). Na umožnenie variability v niektorých pozíciách vzoru sa k danej pozícii môžu pridružiť triedy ekvivalencie, teda skupiny vzorov na tejto pozícii zo syntaktického hľadiska rovnocenných. Veľmi abstraktnou a hrubou triedou ekvivalencie by boli napríklad slová z jedného slovného druhu, napríklad predložky. Pričleňovanie triedy ekvivalencie ku konkrétnejším

vzorom a možnosť použitia malých a kompaktných tried oproti takýmto globálnym triedam umožňuje tesnejšie zomknutú a tým čiastočne aj plauzibilnejšiu reprezentáciu vzorov zoskupujúcich viaceré podobné slovné spojenia. Zložitejšie a aj abstraktnejšie vzory sa dajú získať zoskupovaním podobných jednoduchších a konkrétnejších vzorov. Preto môžu vzory v sebe využívať iné vzory a triedy ekvivalencie vyberať aj medzi vzormi. Takto sa medzi vzormi môžu vytvárať hierarchie konštrukcií od globálnej formy vety k lokálnym inštanciám fráz Vzorom tiež nič nebráni priamo či nepriamo sa odkazovať na seba samých, čím sa poskytuje priestor na reprezentovanie rekurzívnych štruktúr, aké možno vidieť v prirodzenom jazyku.

Tento spôsob, akým sa dajú pomocou konštrukcií skladať zložitejšie jazykové štruktúry reprezentujúce skupiny a postupnosti existujúcich jednoduchších jednotiek je jednoduchým a prijateľným spôsobom, ako vytvárať z konkrétnych konštrukcií viet korpusu zovšeobecňujúcu a pritom plauzibilnú konštrukčnú gramatiku.

Získavanie vzorov

Princípom metódy, ako získavať konštrukcie je postupná redukcia redundancie v dátach. Predpokladom tejto metódy je, že redundancia dát je odrazom použitých signifikantných vzorov inštancovaných vo vetách korpusu. Sledovanou redundanciou je opakovanie zhodných alebo len čiastočne sa líšiacich postupností konštrukcií, ktoré naznačujú svoje nenáhodné usporiadanie a rovnocenné postavenie vyskytujúcich sa odlišných prvkov. Ak sú náznaky o signifikantnosti vyčlenenia takýchto postupností, vytvorí sa zo spoločných častí tejto skupiny nový vzor. Prípadné odlišnosti v nejakom mieste postupnosti sa unifikujú pomocou triedy ekvivalencie pridruženej k novovzniknutému vzoru.

Obmedzenie sa na relatívne malo početné a o to užšie zviazané postupnosti umožní získavať vzory, ktoré sú viac závislé na svojom kontexte, a teda – sú v zhode s očakávaniami – aplikovateľné len za vybraných podmienok a majú tak vyššiu šancu obsiahnuť v sebe skutočnú pravidelnosť v syntaxi a konkrétnejší či praktickejší význam.

Signifikantnosť vzoru, ktorý subsumuje danú skupinu postupností je určovaný prostriedkami pravdepodobnostnej inferencie. V rámci nej sa posudzujú vzťahy medzi k–gramami konštrukcií vo vetách korpusu. Tým sa ponášajú na prístupy využívajúce Markovove modely (premenlivého rádu – VOMM). Zásadná odlišnosť ale spočíva v tom, že trénovanie VOMM je hľadaním parametrov pre vopred danú štruktúru

gramatiky, kým Adios hľadá práve túto štruktúru (paradoxne závisiac od niekoľkých vopred zadaných riadiacich parametrov). Maximálna veľkosť okna (dĺžka postupnosti), ktorú postihuje VOM je vopred limitovaná aj vďaka komplexite príslušných algoritmov. Adios pracuje na skupinách postupností konštituentov danej (možno obmedzenej) dĺžky, prvky týchto postupností ale môžu zastupovať postupnosti konštituentov väčších a dokonca premenlivých dĺžok, čím sa umožní usudzovať o usporiadaní konštituentov, ktoré sú v pôvodných vetách od seba ľubovoľne vzdialené a zoskupovaní viet rôznych a tiež neobmedzených dĺžok. Menšia zložitosť algoritmu spočíva v lokálnosti spomínanej pravdepodobnostnej inferencie: počítané vlastnosti postupností sa určujú zo štatistík obmedzených na konštrukcie z týchto postupností, ostatné vety korpusu výsledok ohodnotenia signifikantnosti vzoru neovplyvňujú.

Vzniknuté vzory tvoria (bez rekurzie) hierarchiu štruktúr vyťaženú z danej vzorky jazyka bez predchádzajúceho určenia možností a obmedzení na gramatiku v dátach ani v modeli. Tieto štruktúry obsahujú štatisticky a potenciálne sémanticky významné pravidelnosti v korpuse. Táto hierarchia ako aj samotné štruktúry vzniknú a usporiadajú sa samostatne, teda bez vonkajšej kontroly učiteľa. Preto možno vzniknutú gramatiku považovať za emergentnú vlastnosť redukcie redundancie viet a v tomto zmysle aj hovoriť o načrtnutom prístupe ako o akvizícii gramatiky pomocou samoorganizácie.

3.2 Popis algoritmu a získanej gramatiky

Model Adios pozostáva z dvoch komponentov: statickú časť tvorí reprezentačná dátová štruktúra (RDS), čo je orientovaný multigraf obsahujúci vstupný korpus viet, ako aj vytváranú gramatiku; dynamickou zložkou je algoritmus akvizície vzorov, ktorý vyhľadáva v RDS grafe významné vzory a podľa nich túto štruktúru upravuje.

Vstupom pre algoritmus sú čisté (neanotované) vety skúmaného jazyka zoskupené do korpusu. V prípravnej fáze algoritmu sa vstupný text rozdelí na najjednoduchšie zmysluplné jednotky (tokeny), s ktorými sa mieni ďalej pracovať. Tak sa napríklad v angličtine oddelí z konca každého slova morféma „ed“, ktorá môže označovať minulý čas pravidelného slovesa, či už opodstatnene ako napr. zo slovesa walked, ale aj nesprávne ako zo slova bed. Odhaliť rozdiel vo význame medzi týmito dvoma výskytami morfémy napríklad jej vyžadovaním vo vzoroch zodpovedajúcich minulému času oproti pevným spojeniam

vhodných výskytov dvojíc tokenov „b“ a „ed“ je súčasťou úlohy pre algoritmus. V zásade môžu byť týmito najmenšími jednotkami aj jednotlivé grafémy či fonémy vyskytujúce sa v korpuse.

Reprezentačná štruktúra

Z takto rozdeleného korpusu sa následne vybuduje počiatočná RDS štruktúra. Každý token, ktorý vznikne pri rozdeľovaní korpusu sa reprezentuje ako jeden uzol RDS grafu. Okrem týchto tokenov sa do grafu pridajú ešte dva špeciálne uzly – begin a end, ktoré označujú začiatok resp. koniec vety. Zmyslom týchto dvoch uzlov je okrem uľahčenia prístupu k vetám tým, že má každá veta ten istý počiatočný uzol aj ich funkcia ako kontextu pre krajné tokeny viet – umožňuje to získavať vzory, ktorých použiteľnosť sa viaže na začiatok či koniec vety; taktiež ako dummy uzly pomôžu vyhladiť algoritmus a počítané vzorce od možných okrajových podmienok.

Každá veta korpusu sa v grafe reprezentuje nezávislou jedinečnou orientovanou cestou od uzla begin k uzlu end, postupne prechádzajúcou uzlami pre jednotlivé tokeny ako vo vete. Označenie hrán grafu jednoznačne identifikuje vetu, ku ktorej patrí, ako aj jej umiestnenie v tejto vete (teda poradové číslo tokenu, z ktorého uzla vychádza, resp. do ktorého vchádza). Takýmto spôsobom sa reprezentuje celý korpus dát do ucelenej štruktúry, ktorá obsahuje všetku informáciu obsiahnutú v texte korpusu a navyše pri sebe uchováva všetky výskyty každého k-gramu všetkých dĺžok. Pretože náhodné cesty grafu tvorené hranami z rôznych viet v sebe nenesú zmysluplné informácie o vzorke dát, uvažuje sa v ďalšom texte len o cestách s nadväzujúcim označením hrán.

Vzory, ktoré tvoria hľadanú gramatiku sa v RDS grafe odrazia ako tendencie mnohých jednotlivých viet/ciest zdieľať spoločné podcesty a vytvárať tak v spletenci grafu zväzky. Prítomnosť signifikantných vzorov sa teda deteguje dostatočne silnými (teda početne zastúpenými), previazanými a oddelenými zväzkami (teda vedúcimi rovnakými dráhami, pričom cesty mimo zväzku k nemu jednoznačne neprislúchajú).

Algoritmus získavania vzorov

Algoritmus získavania vzorov opakovane vyhľadáva práve takéto zväzky. Podľa zvolených kritérií vyberá množiny súvislých podciest zložených z nadväzujúcich hrán vždy tej istej vety. Podcesty v rámci každej tejto množiny majú rovnakú dĺžku a spoločný kontext, teda neprázdny spoločný prefix aj sufix. Rozdiely medzi jednotlivými

podcestami sa môžu vyskytovať nanajvýš na jednom mieste, spoločnom pre všetky prvky zväzku.

Viacero vybraných zväzkov medzi sebou súťaží o vytvorenie vzoru pre prebiehajúcu iteráciu. Tento kandidátsky vzor je zreťazením spoločných častí postupností v zväzku a prípadnej vhodnej triede ekvivalencie zahladzujúcej odlišnosti medzi cestami.

Kandidátske zväzky ciest

Spôsob, akým sa vyberajú relevantné kandidátske zväzky môže základne spočívať napríklad v postupnom výbere každej dostatočne dlhej cesty grafu a niektorého (postupne každého) jej prvku. Zväzok sa predlžuje o nasledujúci uzol z vybranej podcesty, ak je prechod k tomuto uzlu spomedzi ciest v doterajšom zväzku dostatočne pravdepodobný, napríklad pravdepodobnejší ako predchádzajúci prechod. Podobné kritérium platí aj pre predchádzajúce prvky vybranej postupnosti, teda pravdepodobnosť, že budú predchádzať cestám v doteraz vybudovanom zväzku. (viď obrázok 2 v (Edelman et al. 2004))

Medzi kandidátske zväzky sa dostanú len tie s dostatočným zastúpením v korpuse, teda ktorých počet podciest je väčší ako nejaká pevná hranica, ktorá predstavuje jeden z voľných parametrov modelu, pričom nie je nutným cieľom experimentovania s dátami snažiť sa tento parameter nejakým spôsobom optimalizovať, hoci jeho voľba môže ovplyvniť napríklad prítomnosť idiómov (či konštrukcií s vlastnosťami idiómov, ktoré vyplývajú z problému nízkej početnosti javov v korpuse) v konštrukciách výslednej gramatiky (teda ich odfiltrovať v prípade, že sú pri skúmaní či použití gramatiky nezaujímavé alebo zbytočné).

Kandidátske zväzky medzi sebou súťažia svojím ohodnotením lokálnou hodnotiacou funkciou, teda funkciou, ktorá by mala vyzdvihovať zväzky postihujúce vzory žiadúce na zaradenie do gramatiky. Hodnota signifikantnosti sa taktiež porovnáva s globálnou hodnotou minimálnej akceptovateľnosti, ktorú musí dosiahnuť (aspoň) víťazný signifikantný vzor. Toto je ďalší voľný parameter modelu, ktorého hodnota určuje, aké dobré musia byť pri danej miere všetky vyťažené vzory, zvyšovanie tejto hodnoty vedie k vyberaniu vzorov s dobrou podporou v korpuse, zaplateným nižším pokrytím jazyka získanou gramatikou. Výber tejto hodnoty teda závisí od sledovaných cieľov požadovaných od použitia algoritmu.

Ohodnotenie signifikantnosti zväzku

Autori modelu Adios vo svojich prácach uvádzajú dve do veľkej miery si podobné alternatívy takéhoto ohodnocovania zohľadňujúce rôzne odtiene pohľadov a motivácií pre rozhodovanie o vhodnosti a zmysluplnosti daného vzoru. Tieto rôzne pohľady následne vyžadujú aj alternatívne, vzájomne zameniteľné verzie výpočtovej časti algoritmu.

Prvou alternatívou je v ohodnotení zväzku zohľadniť jeho dĺžku v porovnaní s očakávanou priemernou dĺžkou vzorov (L) v hľadanej gramatike, čo je ďalší parameter algoritmu, tentoraz s výrazným vplyvom na výsledný tvar gramatiky. Preferencia krátkych vzorov vedie k zväzkom s výraznejšou podporou v korpuse, na druhej strane znamená kratšia cesta menší kontext a s tým spojené zvýšené riziko prílišného zovšeobecnenia.

Druhou zložkou hodnotenia, vyvažujúcou prvú je miera určujúca kompaktnosť daného zväzku a podporu preň, teda odhaduje, či je zväzok dostatočne zastúpený v korpuse a či postupnosti zo zväzku nemohli vzniknúť náhodnou interakciou iných (skutočných, v zmysle lepších) vzorov, napríklad poskladaním pravdepodobných (vysoko zastúpených) tokenov a bigramov.

Skóre celého zväzku vznikne sčítaním jednotlivých skóre pre jeho cesty C_i podľa vzorca:

$$S(C_i) = e^{-(L/k)^2} \cdot P^{(k)}(C_i) \log \frac{P^{(k)}(C_i)}{P^{(2)}(C_i)}$$

$$P^{(k)}(C_i) = P(c_1)\,P(c_2|c_1)\,P(c_3|c_1 \to c_2) \cdot \ldots \cdot P(c_k|c_1 \to c_2 \to \ldots \to c_{k-1})$$

$$P^{(2)}(C_i) = P(c_1)\,P(c_2|c_1)\,P(c_3|c_2) \cdot \ldots \cdot P(c_k|c_{k-1})$$

$$kde\ C_i = c_1 \to c_2 \to \ldots \to c_k$$

$P^{(k)}$ je pravdepodobnosť výskytu daného k–gramu (celej postupnosti) v korpuse, zatiaľčo $P^{(2)}$ je súčin pravdepodobností jednotlivých bigramov, ktorých zreťazením táto postupnosť vznikne.

Všetky hodnoty pravdepodobností sa odhadnú zo zastúpenia príslušných prvkov a postupností v aktuálnom grafe pomocou Maximum–Likelihood odhadov bez vyhladzovania: pravdepodobnosť $P(c_j)$ sa odhadne ako pomer početnosti prvku c_j v grafe, teda počtu hrán, ktoré vychádzajú z príslušného uzla (ktorý sa zrejme rovná počtu vchádzajúcich hrán) a sume týchto početností pre všetky uzly grafu. Za pravdepodobnosť každého bigramu sa dosadí hodnota relatívneho zastúpenia hrán vedúcich do druhého člena dvojice medzi hranami

vychádzajúcimi z jej prvého člena. Pravdepodobnosťou nadväznosti ďalšieho člena na danú postupnosť je zase pomer počtu postupností pokračujúcich do tohto člena a počtu výskytov tejto postupnosti.

Ako vidno, všetky hodnoty pravdepodobností sú lokálne obsiahnuté v aktuálnom RDS grafe, pričom sa počty sledovaných hrán dajú jednoducho získať priamo počas zoskupovania použitých ciest (čo je zrejmou nutnou súčasťou hľadania kandidátskych zväzkov). Použitá globálna hodnota celkovej veľkosti grafu sa môže vypočítať pri jeho konštrukcii ešte pred prvou iteráciou; následne sa aktuálna hodnota upravuje podľa zmien vykonaných na grafe. Ostatne, pre danú iteráciu je táto hodnota konštantná a vyskytuje sa pre každú cestu na rovnakom mieste, a zo vzorca vyplýva, že neovplyvňuje poradie kandidátskych zväzkov podľa daného ohodnotenia signifikantnosti. Preto sa môže aj vynechať, aj keď skutočná hodnota skóre môže ovplyvniť globálnu prijateľnosť vzoru. Spomínaná lokálnosť je rozhodne veľkou výhodou modelu Adios, keďže umožňuje pracovať aj s obrovskými korpusmi, pre ktoré by bolo potenciálne prehľadávanie celého grafu zložité až prakticky neúnosné. Z tohto dôvodu sa tiež používajú len jednoduché odhady pravdepodobnosti, teda bez vyhladzovania.

Druhou alternatívou merania významnosti kandidátskych vzorov je takpovediac dvojfázové súťaženie: v prvej fáze sa vyberie víťazný zväzok dlhší ako pri prvej alternatíve – signifikantná cesta a v druhej sa z neho vyberie najlepší kratší zväzok – signifikantný vzor.

Signifikantná cesta sa vyberie v zásade zhodným spôsobom ako pri prvom prístupe. Zo vzorca pre ohodnotenie cesty, a tak aj z celého modelu vypadne parameter očakávanej dĺžky vzoru. Takto sa vzorec pre celkové ohodnotenie zväzku (suma ohodnotení ciest v ňom obsiahnutých) zmení na vzorec pre Kullback–Leiblerovu divergenciu, čo objasňuje použitie daných vzťahov.

Pri danej interpretácii sa medzi sebou porovnajú dve hypotézy: nulová hypotéza, ktorú sa snažíme vyvrátiť, je hypotéza o tom, že cesty v kandidátskom zväzku sú náhodným pospájaním často sa vyskytujúcich bigramov. Alternatívna hypotéza hovorí, že cesta je v celej svojej dĺžke tvorená prvkami závislými od všetkých predchádzajúcich, čiže s nimi previazanými silnou väzbou, ktorá je odrazom existencie vzoru pre takéto postupnosti v gramatike, ktorá daný korpus vygenerovala.

Matematickým vyjadrením týchto dvoch hypotéz vo forme ich vierohodnosti v prítomnosti zozbieraných dát sú práve členy zlomku zo vzorca. Vypočítaná hodnota skóre bude rásť podľa toho, ako sú cesty zo zväzku lepšie popísateľné pomocou alternatívnej hypotézy. Pri danej hodnote skóre sa dôveryhodnosť v jej výpovednú hodnotu zvyšuje proporcionálne s kvalitou nulovej hypotézy, teda jej schopnosťou byť pravdepodobnejšia ako alternatívna v tých prípadoch, keď alternatívna hypotéza neplatí. Prirodzený jazyk používaný ľuďmi je tvorený postupnosťami symbolov s nenáhodným usporiadaním a nie ich voľnými zoskupeniami (v angličtine používaný termín je „bags of words“). Preto je vhodnejšou nulovou hypotézou práve tá o náhodnom zreťazení postupností ako trebárs hypotéza o náhodnom výskyte jednotlivých prvkov. Jej vierohodnosť by bola vyjadrená súčinom pravdepodobností (teda relatívnych zastúpení) týchto prvkov. Táto vierohodnosť by bola často omnoho nižšia ako použitá, čím by sa dala jednoduchšie prekonať a pustiť tak aj väčší počet menej kvalitných zväzkov. Bigramová hypotéza tiež nie je príliš zložitá na to, aby bola jej hodnota vysoká pre skutočne náhodné (ale pritom jazyk napodobňujúce) dáta.

Na víťaznej signifikantnej ceste sa v druhom kroku vyhľadá konečný signifikantný vzor. Tento je vyjadrený skupinou podciest zväzku (so spoločným umiestnením v ňom, teda skupina rovnako dlhých podciest pôvodných ciest s tým istým odsadením od začiatku) s najväčšou súdržnosťou. Ak e_1 až e_k sú množiny konštituentov na príslušných pozíciách v podcestách zväzku, vypočíta sa stupeň súdržnosti podcesty začínajúcej v e_i a končiacej v e_j podľa vzorca:

$$s_{ij} = P_{ij} \cdot \log \frac{P_{ij}}{P_{i,j-1} P_{i+1,j}}$$

$$P_{ij} = \begin{cases} P(e_j \mid e_i \rightarrow e_{i+1} \rightarrow \ldots \rightarrow e_{j-1}) & \text{ak } i < j \\ P(e_j) & \text{ak } i = j \end{cases}$$

Súdržnosť sa teda takisto zisťuje testovaním hypotéz. Tentokrát proti sebe stoja hypotéza o súdržnosti celej podcesty oproti nulovej hypotéze o tom, že sledovaná podcesta je lepšie opísaná ako prelínajúci sa produkt dvoch o jeden prvok kratších podciest. Oproti predchádzajúcim metódam sa ale posudzujú početnosti prechodov medzi množinami prvkov ciest spoločne a nie pre každú podcestu zvlášť. Hodnoty P_{ij} sa ale stále odhadujú pomocou počtov prvkov nadväzujúcich v niektorej vete korpusu, tieto sa dajú efektívne počítať metódou analogickou s dynamickým programovaním.

Vznik nového vzoru a triedy ekvivalencie

Tvar nájdeného významného vzoru sa zapamätá (nie nutne ako postupnosť uzlov grafu) a jeho použitie sa uchová v RDS grafe rovnako ako použitie jednoduchších konštituentov – pôvodných tokenov. Zo vzoru vznikne nový uzol, cez ktorý sa presmerujú všetky cesty víťazného zväzku. Nový uzol tak prevezme len tie cesty z tých, ktoré v korpuse subsumuje, ktoré prispeli k jeho vzniku. Cesty, ktoré vedú cez uzly zo sekvencie vzoru, ale neboli dostatočne podobné pre výber do zväzku ostanú danou iteráciou nedotknuté, kým informácia o pôvodnom znení presmerovaných ciest sa z grafu vylúči. Zabezpečí sa tak, že sa do vzoru započítajú len jeho výskyty v požadovanom kontexte a už subsumované inštancie zas nebudú priamo ovplyvňovať ďalšie hľadanie, a tým brániť slabším, hoci tiež významným vzorom v súťažení.

Ako bolo spomenuté skôr, na mieste vo vzore, kde sa povolil výskyt rôznych konštituentov sa vzoru priradí trieda ekvivalencie, zahŕňajúca všetky konštituenty považované za vzájomne zameniteľné (len zo syntaktického hľadiska, zrejme nejde o synonymá). Na zvýšenie kompaktnosti reprezentácie je možné pri priraďovaní tried ekvivalencie znovu využiť už existujúce triedy. Dôvodom je možnosť obohatiť nové vzory aj o ďalšie prvky množín, ktoré sú ekvivalentné so súčasnými, ako sa zistilo počas predchádzajúcich iterácií. Takýmto spôsobom sa dá do istej miery brániť voči problému riedkych dát, hoci to prináša riziko zavlečenia nevhodných prvkov do vzoru a oslabenie jeho sily ako kontextovo závislej konštrukcii. Preto je vhodné využívať predošlé triedy ekvivalencie len za vhodných podmienok (závisiacich od cieľov, ktoré má akvizícia dosiahnuť), alebo vytvárať pre každý vzor novú triedu, tak ako v prípade, že žiadna z existujúcich nie je nadmnožinou prvkov, ktoré treba zoskupiť v novom vzore.

Iterácie algoritmu

Algoritmus získavania vzorov pokračuje vo svojej činnosti ďalšou iteráciou. Po čase sa do nových vzorov zaradia aj vzory z predchádzajúcich kôl, čím sa rekurzívne vyťaží ich hľadaná hierarchická štruktúra. Na výber kandidátskych zväzkov môže tiež vplývať kolekcia tried ekvivalencie: konštituenty, ktoré patria do vybranej triedy sa môžu považovať za totožné pre potreby spoločných podpostupností v cestách zväzku, čím sa obohatia možnosti na formy, ako aj štatistická podpora vzorov. Proces sa opakuje dovtedy, kým sa z grafu dajú vyťažiť nové

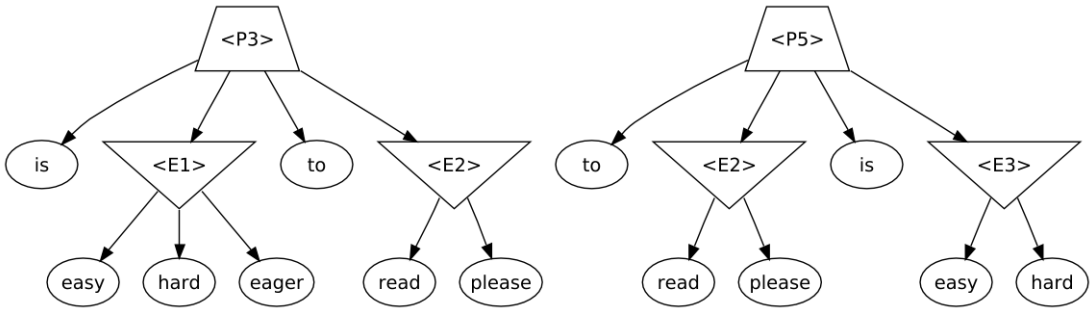
vzory, alebo až kým sa nepodarí vytvoriť dobre podporený vzor na zvolený počet pokusov.

Vo finálnom grafe sú obsiahnuté všetky vety zdrojového korpusu, ktorých niektoré redundantné časti boli zredukované do vzorov. Z tohto pohľadu sa dá konečná RDS štruktúra považovať za kompaktnejšiu reprezentáciu vstupu, ktorá by ako gramatika mala perfektné vybavenie si (recall) trénovacích dát. Pre potreby vyhodnotenia algoritmu sú ale zaujímavé len získané vzory a triedy ekvivalencie, keďže cieľom algoritmu je vložiť práve do nich pravidlá, ktorými sa riadi skúmaný jazyk. Výslednou gramatikou sa teda stane len táto skupina konštituentov (spolu s tými pôvodnými konštituentami – tokenmi, ktoré sú v nich použité). Na zavedenie rekurzie do vzorov sa v prípade potreby môžu vzory medzi sebou porovnať a vhodne zlúčiť či doplniť.

Gramatika ako hierarchia vzorov

Spomínaná hierarchia vzorov sa dá vizualizovať graficky. Vzory gramatiky predstavujú požiadavku na súčasný výskyt skupiny konštituentov v danom poradí, kým triedy ekvivalencie vyžadujú výskyt ľubovoľného jediného z nich. Vizualizácia tak nadobúda tvar lesa usporiadaných a/alebo stromov, kde vzory sú a–uzly a triedy ekvivalencie alebo–uzly so svojimi konštituentami ako synmi (teda sa vzor v lese môže vyskytovať aj viackrát); listy stromov tvoria reťazcové tokeny. Frázy vyjadriteľné vzorom sa tak dajú chápať ako pokrytia stromu v ňom zakoreneného, pričom zrejme z a–uzla vyberáme všetkých synov v určenom poradí a z alebo–syna práve jedného syna.

Táto gramatika sa dá zjavne reprezentovať aj pravidlami bezkontextovej gramatiky v určitom normálnom tvare: vzory a triedy ekvivalencie tvoria množinu neterminálov a tokeny množinu terminálov gramatiky, pričom pravidlo pre každý vzor je jediné, prepisujúce neterminál na sekvenciu terminálov a neterminálov a trieda ekvivalencie má pravidlá prepisujúce ju na každý jeden jej člen, alebo skrátené pravidlo vyberajúce jedinú alternatívu. Bez zavedenia rekurzie je uvedená gramatika konečnej hĺbky, a teda k nej existuje aj regulárna gramatika. Vskutku, tvar vzorov je interpretovateľný aj ako sada regulárnych výrazov. Tieto regulárne výrazy v sebe ale potenciálne zahrňujú všetky úrovne rekurzie skutočne využité v korpuse, preto nemožno takýto druh gramatiky zavrhnúť.



Obrázok 1: Príklad signifikantných vzorov gramatiky. Generalizácia sa dosahuje výberom kombinácií prvkov z triedy ekvivalencií ktoré sa v pôvodnom korpuse potenciálne nevyskytovali. Pre zachovanie gramatickosti viet je potrebné vytvárať triedy ekvivalencie závislé na svojom kontexte. Takýmto spôsobom sa môžu dodržiavať ohraničenia (zhody v gramatických kategóriách) medzi (aj vzdialenými) frázami vo vete alebo sémantická korektnosť celých fráz (napríklad vyhnutie sa nesprávnej fráze „to please is eager" v príklade).

Ak by gramatika tvorila len tie vety, ktoré sa vyskytli vo vstupnom korpuse, nespĺňala by svoj účel. Predstavované vzory ale spoločne majú schopnosť generalizácie aj na nevidené vstupy: superpozíciou dvoch tried ekvivalencie za sebou sa dáva možnosť gramatike voľne kombinovať prvky oboch z nich, keďže tie sa považujú v príslušných miestach za rovnocenné. Počet rôznych takýchto kombinácií je rovný súčinu veľkostí tried oproti počtu rôznych fráz, z ktorých boli tieto triedy zozbierané, ktorý je maximálne rovný ich súčtu. V prípade orientácie akvizície gramatiky na vysoko generalizujúce vzory je preto možné počítať pomer videných a nových fráz v danom vzore obsiahnutých a nevhodné vzory odmietať pomocou ďalšieho globálneho kontrolného parametra.

Testovanie gramatiky

Schopnosti gramatiky postihnúť jazyk v celej jeho šírke sa dajú merať pomocou miery aktivácie, akú dosahujú vzory s vetami jazyka. Komplementárna schopnosť gramatiky generovať dobré vety jazyka je z hľadiska testovania obtiažnejšia. Zlatý štandard pre gramatiku žiadneho prirodzeného jazyka totiž neexistuje, a teda sú výsledky určované buď manuálnym analyzovaním viet, alebo štúdovaním vnútornej reprezentácie gramatiky, v oboch prípadoch však sú hodnotené subjektívnym pohľadom človeka.

Spomínaná miera aktivácie je ale lepšie vyjadriteľná presnými kritériami a to pomocou maximálnej zhody viet so vzormi. Zhoda sa počíta v hierarchii zdola nahor. Na strane terminálov t_i sa vypočíta ich aktivácia jednotlivými tokenmi s_j vety ako hodnota:

$$a_{ij} = P(t_i, s_j) \cdot \log \frac{P(t_i, s_j)}{P(t_i)P(s_j)} \quad ,$$

kde $P(t_i, s_j)$ je pravdepodobnosť kookurencie terminálu a tokenu v nejakej triede ekvivalencie a $P(t_i)$ a $P(s_j)$ sú pravdepodobnosti nezávislého výskytu každého z prvkov v nich. Je to teda porovnanie hypotézy o spolupatričnosti oboch prvkov oproti nezávislosti. Možno poznamenať, že tento vzorec je tiež jadrom formuly pre vzájomnú informáciu (previazanosť) dvoch náhodných premenných.

Pre nové tokeny nevyskytujúce sa v pôvodnom korpuse sa ako aktivácia dosadí defaultná hodnota ϵ (napr. 0,01). Umožní to iný druh generalizácie v gramatike. Aktivácia vzorov obsahujúca tento nový token potom napovedá o jej kategorizácii gramatikou (pridružení k tokenom, na ktorých mieste leží).

Tieto aktivácie sa prenášajú do tried ekvivalencie, kde sa aktivácia maximalizuje, a do vzorov, kde sa aktivácia danou podsekvenciou vety počíta ako priemer aktivácií konštituentov vzoru prislúchajúcimi pod–podsekvenciami. Maximálne dosiahnuté aktivácie vzorov sú hľadanou mierou zhody vety s gramatikou, čím sa dajú považovať za druh posudzovania vety gramatikou z hľadiska neostrej gramatickosti.

3.3 Prehľad výsledkov dosiahnutých autormi modelu

Vzhľadom na to, že medzi hlavné priority štatistického spracovania jazyka patrí snaha ukázať, že je možné naučiť sa prirodzený jazyk bez potreby použitia jazykovo špecifických kognitívnych procesov a tým potenciálne vyvrátiť téorie o vrodených, špeciálne na jazyk orientovaných mentálnych modulov v ľudskom mozgu a z nich vyplývajúcu teóriu univerzálnej gramatiky, je dôležitou požiadavkou na modely z tohto smeru všeobecnosť použitých princípov a algoritmov.

Do modelu Adios zjavne neboli vložené nežiadúce pomôcky, algoritmus samostatne vyhľadáva gramatiku jazyka bez predošlých obmedzení na ňu a bez lingvistickej informácie. Širšia použiteľnosť prístupu na vyhľadávanie signifikantných vzorov v dátach sa predviedla jeho využitím autormi mimo domény spracovania prirodzeného jazyka, a to pri funkčnej klasifikácii enzýmov.

Klasifikácia enzýmov

Enzýmy sú druhom bielkovín napomáhajúcim biochemickým procesom v organizmoch. Sú to zložité molekuly tvorené reťazcami dvadsiatich aminokyselín, pričom párová podobnosť medzi enzýmami je indikátorom podobnosti vo funkčnosti enzýmu. Táto funkčnosť sa kategorizuje do štvorúrovňového kódu, kde štvrtá úroveň opisuje presnú funkciu enzýmu a prvé tri enzýmy rôzne (nezávisle) zoskupujú.

V popísanom experimente (Kunik et al. 2004) bol na korpus tvorený bielkovinami použitý zjednodušený vyhľadávací modul Adiosu (nazvaný MEX) na detekciu motívov, teda kompaktných sekvencií aminokyselín významne sa podieľajúcich na funkčnosti. Vytvorený priestor motívov bol potom použitý ako vstupný priestor pre SVM (Support Vector Machine) klasifikátor, ako pri iných riešeniach problému. Dlhšie získané motívy sa ukázali byť dobrými indikátormi významu daného enzýmu (čo je dobrým znamením pre sémantickú plauzibilitu vzorov produkovaných v jazyku), keď prekonali iné modely pri klasifikácii tretej úrovne kódu pri porovnateľnom výkone na druhej. Iné modely pritom neboli celkom bez učiteľa, a ďalšie sa opierali aj o fyzikálne a chemické odlišnosti aminokyselín.

Testovanie pomocou umelých gramatík

Vlastnosti a schopnosti algoritmu Adios zachytávať štruktúru jazyka boli najskôr testované na korpusoch generovaných bezkontextovými gramatikami modelujúcimi jednoduché triedy viet anglického jazyka. Sledovaná bola presnosť, a schopnosť generalizácie v závislosti na veľkosti vstupného korpusu. Medzi výsledky prezentované z týchto experimentov patrili:

- Na konečnej gramatike (generujúcej 2016 anglických oznamovacích viet a podmetových fráz) sa dosiahlo 95% pokrytie po poskytnutí 20 % viet do korpusu; pri 10 % to bola asi štvrtina.
- Z bezkontextovej gramatiky tvaru 28 vzorov Adiosu do tretej úrovne hierarchie sa vyprodukovalo 400 viet, na ktorých sa natrénoval Adios. Prienik generovaných jazykov sa testoval vyprodukovaním 15 mil. viet, z ktorých bolo 3,6 mil. pôvodnej resp. 1,9 mil. viet novej gramatiky rôznych. Pri prijatí 95 % zhody medzi dvoma vetami z korpusu za úspech sa dosiahol podiel naučených viet v cieľových 97 % a opačný podiel 53 %.

- Na bezkontextovej gramatike (generujúcej asi 160 mil. viet) sa trénoval model na korpusoch do veľkosti 6400. Potom sa na 10000 testovaných naučených vetách dosahovala úspešnosť do 90 %.

Použitím prvého spôsobu vyhodnocovania signifikantnosti vzoru (ADIOS1) uvedeného v kapitole 3.2 na reálnych korpusoch dát sa nedosahovali globálne uspokojivé výsledky gramatickosti viet. Model závisel od voľby očakávanej dĺžky vzoru, čo pri jej menších hodnotách nemali zväzky dostatok kontextu na bránenie prílišnému zovšeobecneniu, na dlhšie vzory nebola dostatočná podpora. Preto aj vznikla druhá predstavená, revidovaná verzia algoritmu ADIOS2.

Testovanie na prirodzenom jazyku

Na experimentovanie v podmienkach reálneho jazyka autorom slúžili vety rodičov určené deťom zo zbierky korpusu CHILDES. ADIOS1 sa trénoval na 9665 vetách, pričom vyprodukoval 1062 vzorov s 775 triedami ekvivalencie. Práca algoritmu sa prejavila na reprezentácii korpusu v RDS grafe, ktorá sa zredukovala z priemerných 6,70 konštituentov na vetu na 2,18 po natrénovaní. Frázy generované vzormi, ktoré sa nenachádzali v korpuse, poukazovali na schopnosť algoritmu nachádzať vzory s adekvátnymi triedami ekvivalencie úspešne generalizujúce vety na nové gramaticky správne vety, avšak nie vždy aj s celkom zmysluplným významom či úspešnou zhodou, napríklad osoby a čísla zámena ku slovu.

ADIOS2 bol trénovaný na omnoho väčšom korpuse 300 000 viet. Počas dvojtýždňového spracovávania našiel 3400 vzorov a 3200 tried ekvivalencie. Počas trénovania boli vyťažené vzory testované pomocou testu používaného na zistenie úrovne angličtiny u žiakov stredných škôl vo Švédsku. Test pozostával zo 100 viet s vynechaným jedným slovom a troma možnosťami na jeho doplnenie.

Testovanie prebiehalo tak, že sa zisťovali aktivácie aktuálnej sady vzorov spôsobené každou z troch možných viet. Možnosť s najlepším skóre sa zvolila ako odpoveď, v prípade zhody (resp. nedostatku zhody so vzormi) medzi možnosťami sa neodpovedalo. S rastúcim počtom iterácií sa postupne zvyšoval podiel otázok, na ktoré bola gramatika schopná odpovedať. Vo finálnej fáze dosahoval polovicu viet. Úspešnosť uskutočnených odpovedí sa počas každého testu pohybovala okolo 60 %, čím podľa vyhodnotenia testu dosiahol na zodpovedaných otázkach úroveň „pokročilý."

4 Model AMIGOS[1]

Model Adios zrekapitulovaný v predchádzajúcej kapitole je vskutku inšpiratívnym výtvorom spájajúcim moderné prístupy lingvistov k reprezentácii gramatiky jazyka s výpočtovo, kognitívne a biologicky realistickejším štatistickým pohľadom na jeho akvizíciu. Jeho inovatívne využitie grafovej reprezentácie vstupných sekvencií dát umožňuje efektívny prístup k informácii o podobnostiach medzi jednotlivými vetami. Zároveň sú navrhnuté postupy dostatočne všeobecné, aby boli aplikovateľné aj mimo pôvodnej domény jazyka.

Úspešnosť opísaná prezentovanými výsledkami nasvedčuje o užitočnosti ďalšieho skúmania modelu. Ako mnohé iné výskumy výpočtovej lingvistiky sa aj autori Adiosu aj z praktických dôvodov zamerali pri experimentovaní na anglický jazyk. Odhliadnuc od toho, že je snáď pre každý prirodzený jazyk jeho neobmedzené algoritmické využívanie v počítačoch v súčasnosti neprekonateľným problémom, je známym faktom aj to, že angličtina je oproti iným jazykom v mnohých ohľadoch jednoduchšia. Preto je vždy vhodné preveriť získané poznatky aj na zložitejších problémoch a odhaliť prípadné nedostatky.

Slovenský jazyk je objektívne príkladom zložitejšieho problému ako angličtina. Kým angličtina sa vyznačuje pevným usporiadaním fráz, kde umiestnenie často veľmi mnohoznačného slova vo vete je rozhodujúce pre určenie jeho typu, slovenčina umožňuje viacero alternatívnych usporiadaní podfráz s minimálnymi zmenami odtieňa významu. Medzi hlavné nepravopisné problémy, ktoré mávajú aspoň ľudia učiaci sa slovenčinu ako ďalší jazyk patrí bohatosť tvaroslovia pri potrebe zachovania zhody gramatických kategórií medzi viacerými slovami vety (hoci existujú z tohto hľadiska výrazne zložitejšie jazyky, napr. fínčina).

Z týchto dôvodov a tiež z dôvodu prístupnosti vhodného korpusu slovenského jazyka zozbieraného do Slovenského národného korpusu sa stalo preverenie myšlienok Adiosu na slovenskom jazyku predmetom tejto práce. Vzhľadom na neprístupnosť neobmedzene využiteľnej implementácie ani dôkladného opisu pôvodného modelu, ako aj určité

1 Označenie modelu predstaveného v tejto kapitole vzniklo hlavne z nutnosti nejako nazvať implementačný projekt vo vývojovom prostredí. Vzhľadom na významné zmeny oproti Adiosu vykonané pre potreby tejto práce nebolo použitie rovnakého názvu vhodnou alternatívou, preto bolo vybrané podobné slovo ako náhrada. Amigos môže tiež slúžiť ako akronym pre zmeny charakterizujúcu frázu Adios' morphologically–intensive–grammar–oriented specialisation. Iné názvy, napr. Muchachos podobne dobré využitie neposkytujú.

námietky voči nemu sa v rámci práce na princípoch Adiosu vyvinul nový model, Amigos, pomocou ktorého sa tieto princípy vyhodnocujú. Prezentovaný prístup vďaka viacerým zmenám oproti Adiosu možno nemá priamu výpovednú hodnotu o pôvodnom prístupe, avšak autor tejto práce nepredpokladá principiálne rozdiely medzi globálnym správaním oboch sledovaných modelov.

4.1 Aplikácia ADIOSU a špecifiká slovenčiny

Prvým praktickým a čiastočne aj teoretickým problémom pri aplikácii Adiosu na korpus slovenského jazyka tkvie už v jeho prípravnej fáze. Táto spočíva vo vybudovaní iniciálneho RDS grafu z viet korpusu rozdelených na minimálne tokeny jazyka, čo zahŕňa napríklad oddelenie význačných morfém od každého slova korpusu, v ktorom sa nachádzajú. Tu sa nachádza istý rozpor s princípmi, na ktorých je Adios vybudovaný.

Jedným z týchto princípov je princíp samostatného učenia sa bez učiteľa, teda použitie hrubého („raw"), neanotovaného korpusu. Použitie morfém s význačným postavením v použitom jazyku sa však dá považovať za určitý spôsob anotácie. Dôvodom pre zavedenie toho princípu bol cieľ autorov vyvinúť systém, ktorý hľadá štruktúru, teda gramatiku využitú v jemu prezentovanom jazyku, miesto jej predurčenia, čím by systém učenia slúžil skôr na potvrdzovanie predpokladov o gramatike zo strany experimentátora. Informovanie systému o tom, ktoré jednotky sekvencie sú dôležitou súčasťou danej gramatiky ich samostatným vydelením zo slov, čím sa stanú pre algoritmus vhodným vodítkom na zistenie podobnosti fráz, je zrejme takýmto predurčovaním, tým skôr, že pravdepodobne na úrovni morfém nie je veľký priestor, aby boli tieto predpoklady odtrhnuté od reality. Túto principiálnu výhradu voči navrhovanej praxi neoslabí ani fakt, že sa dané morfémy oddelia aj v ich výskytoch, kde neboli v slove použité kvôli ich gramatickému významu (porovnaj príklad s walked a bed v kapitole 3.2).

Ďalším z cieľov projektu bola snaha o prijateľnosť riešenia ako známky o principiálnej naučiteľnosti prirodzeného jazyka. To so sebou prinášalo ešte väčšiu potrebu nezávislosti na konkrétnom jazyku, ktorú dané predspracovanie porušuje, keďže je z tohto pohľadu (aj deklarovanou) prvou fázou a teda súčasťou samotného algoritmu. Tým pádom sa danej inštancii implementácie nemôže podsunúť korpus ľubovoľného jazyka, aby z neho vyťažila príslušnú gramatiku, čo je zjavnou schopnosťou učenia sa u detí.

Praktickejším problémom z pohľadu použitia algoritmu v tejto práci je bohatosť morfológie slovenčiny, keďže existuje množstvo rôznych tvarov každého slova ohybného slovného druhu, pričom veľká pravidelnosť mnohých z nich v základných vzoroch je vyvažovaná výnimkami a nepravidelnosťami pri spájaní kmeňa slova s príponou a podobne. Aj v prípade prístupnosti vhodného zoznamu morfém daného jazyka (ktorý by pre niektoré jazyky bol veľmi obsažný) by sa tým neriešili predchádzajúce výhrady. Aj preto sa táto práca podujala vzniknutý problém riešiť modifikáciou modelu.

Oba druhy problémov (teda teoretické aj praktické) autori Adiosu samozrejme predvídali. Riešenie spočíva v použití samotných znakov (grafém či morfém) ako základných stavebných jednotiek a uzlov grafu, keďže ignorovanie morfológie jazykov a použitie celých slov miesto toho by viedlo k nepoužiteľne nesystematickým gramatikám vzhľadom na jej využitie ľuďmi pri tvorbe nových tvarov slov (hoci táto tvorba nemusí nutne byť priamou súčasťou pravidiel tvorby komunikovaných viet teda gramatiky, ale prácou vyššieho stupienka myslenia, ktorého využívaním aj pri analyzovaní a konštruovaní gramatík sa tento môže do gramatiky mimovoľne zahrnúť).

Návrh na delenie po morfémach totiž nie je nutnou podmienkou fungovania algoritmu, ktorý principiálne dokáže pracovať aj so znakmi. Dôvodom pre navrhnutý postup bolo skôr ľahšie zrozumiteľné predvedenie práce algoritmu na zaujímavejšej časti procesu získavania gramatiky. Prvá fáza preto preskočila menej intuitívne časti procesu (samo)organizácie dát v korpuse. Ďalším dôvodom môže byť možná potreba omnoho väčšej (až technicky ťažko zvládnuteľnej) veľkosti korpusu a počtu potrebných iterácií algoritmu na dosiahnutie porovnateľných výsledkov. Toto nie je problémom pri prirovnávaní s učením sa detí, kde možných a paralelne spracovateľných (skôr medzi sebou neodlíšiteľných) iterácií je zrejme obrovské množstvo).

Problémy spracovania korpusu po znakoch

Vo fáze počiatočného spájania znakov do slabík a následne slov by algoritmus musel pracovať veľmi obozretne: slabinou Adiosu je jeho monotónnosť, teda nemožnosť revidovať svoje rozhodnutia a vzory vykonané v predchádzajúcich iteráciách. Každý vzor totiž ireverzibilne nahradí subsumované podcesty grafu, čím sa stratí možnosť alternatívneho zaradenia konštituentov do vzorov. Určitým uvoľnením

tohto problému v biologickom učení sa jazyka[2] je to, že dieťa dostáva svoj vstupný korpus po častiach: hoci (jeho mozog, viď poznámka pod čiarou 2) spraví chybu pri vytváraní vzoru, má možnosť vytvoriť aj iné vzory na novej časti vstupu obsahujúcej aj podcesty, ktoré by boli počas vytvárania vzoru nahradené. V našom modeli by sa to dalo simulovať spracovávaním korpusu po častiach.

Veľmi nevhodným, aj keď veľmi pravdepodobným správaním algoritmu by bolo aj vytváranie vzorov písmen prekonávajúcich hranice slov (prípadne obsahoval medzery, ak by ich aj vstup obsahoval) so súčasným lámaním susedných slov. Tento jav môže byť spôsobený morfémou (ktorá má zrejme nadštandardné zastúpenie vo vetách korpusu) spojenou s častým prefixom (či napríklad písmenom p v slovenčine). Podobný efekt sa dá pozorovať aj v inej práci (Wolff, 1988) zaoberajúcej sa indukciou gramatiky z písmen.

Opatrnosť spájania znakov by sa pri behu algoritmu musela prejaviť veľmi pomalým postupom opakovaného spájania malých skupiniek inštancií znakov (hrán grafu), čím by vzniklo nepreberné (z pohľadu skúmania výsledkov algoritmu neprehľadné) množstvo vzorov s totožnou sekvenciou. Potenciálne pripustiť triedy ekvivalencie v tomto štádiu spracovania korpusu by bolo tiež považlivé (a použitie parametra algoritmu filtrujúceho dobré generalizácie vylúčené).

Psychologickým problémom takýchto vzorov by bolo priradiť im prijateľnú zložku významu, ktorú by mali vzory tvorené Adiosom niesť. Deklarovaná totiž bola zamýšľaná a želaná príbuznosť medzi získanými vzormi a konštrukciami v ich ponímaní pri konštrukčných gramatikách. Prípustné konštrukcie tam totiž priradenie (hoci aj neúplného) významu vyžadujú.

2 V spojení s prezentovaným prístupom sa vynára otázka jeho biologickej akceptovateľnosti. Kým biologicky sa jazyk spracováva, organizuje a tvorí neurónmi prostredníctvom subsymbolových výpočtov a možnosť (vedomého?) symbolového spracovania jazyka prichádza u detí pravdepodobne nie skôr ako sa jazyk začne vyvíjať, algoritmus Adios pracuje na diskrétnej, symbolovej a abstraktnej (navyše algebraickej!) štruktúre – RDS grafe, kde počíta a preraďuje hrany a dotvára uzly.

Použitú reprezentáciu treba zrejme chápať ako abstrakciu procesov prebiehajúcich v mozgu. V súlade so súčasnými neurovednými teóriami možno potom chápať jednotlivé konštituenty, teda uzly grafu ako reprezentácie súčasnej aktivácie nejakej skupiny neurónov distribuovane reprezentujúcich tieto konštituenty v mozgu. Hranami grafu sa reprezentuje časová následnosť (aktivácie) konštituentov vo vetách. Počet hrán vedúcich jedným smerom v grafe je vyjadrením častého opakovania nejakej postupnosti vzorcov aktivácií. Vzor je následkom

Iným súvisiacim problémom je významná črta algoritmu: ako kandidátske zväzky sa uvažujú len skupiny ciest rovnakej dĺžky. Vzhľadom na to, že očakávať dostatočnú podporu pre spojenie kmeňa každého (alebo veľkej časti) slova do jediného konštituentu by bolo príliš optimistické, zlučovať do tried ekvivalencie by sa reálne mohli len slová podobných dĺžok, čo nie je opodstatnené ani (jednoducho a tiež psycholingvisticky) obhájiteľné.

Z týchto dôvodov vzniká (pri prakticky použiteľnej realizácii; analógia z poznámky 2 by asi dokázala podobné výhrady prekonať) potreba zachovať informáciu o celistvosti tokenov (tentoraz celých slov) ako aj umožniť algoritmu nazrieť do ich vnútornej štruktúry (aby túto štruktúru mohol *nájsť*). Existencia prijateľného kompromisu a z neho sa odvíjajúcich zásadných zmien modelu (a upresnení ďalších výhrad) tvorí hlavný príspevok tejto práce v druhej časti tejto kapitoly.

4.2 Navrhované riešenie a následné rozhodnutia

Jednoduchým a pochopiteľným riešením je využitie rozdelenia viet po znakoch doplnené o priradenie explicitného, význačného významu medzere, respektíve hranici medzi tokenmi a jeho priame využitie v algoritme[3]. Pomôcka, ktorá sa takto algoritmu podsunie nie je taká problematická ako návrhy v predchádzajúcom modeli.

Z teoretického pohľadu nie je medzera rôzna medzi jazykmi (ako napríklad morféma „ed“, ktorú využíval model Adios navyše k jeho využitiu hranice medzi slovami), ani nie je priamym vkladom nejakej konkrétnej gramatickej teórie – každá uvažuje o nejakých hraniciach medzi frázami. Predpoklad o usporiadaní vnemov do blokov tiež nie je nijako jazykovo špecifický.

---

previazanosti medzi aktiváciami neurónov v postupnosti získaný Hebbovským učením, čiže tiež (skoro) súčasnou aktiváciou skupiny neurónov, ktorá sa reprezentuje uzlom grafu. Presmerovanie dráh v grafe do uzla vzoru je vysvetliteľné ako spustenie príslušnej série aktivácií zabezpečené silou prepojenia medzi jednotlivými asociovanými sadami neurónov. Trieda ekvivalencie je jednotným vyjadrením väčšej skupiny vzorcov aktivácií asociovanej na určitej pozícii v sérii. Gramatika vznikne emergenciou z veľkého množstva asociácií.

Hoci je načrtnutá analógia zaujímavým (až prekvapivým) pohľadom autora práce na stupeň plauzibility modelu Adios, nie sú aplikácie neurónových sietí napriek jej titulu predmetom súčasnej diplomovej práce.

3 Autor nemá podrobný prehľad o rôznorodosti prirodzených jazykov. Uvedomuje si ale, že sa pri odpútavaní modelu Adios od angličtiny nemusel dostať napríklad z okruhu indoeurópskych jazykov, kde možno skutočne vravieť o hranici medzi

Z praktického hľadiska tiež nie je hranica medzi tokenmi žiadnym problémom. Medzery medzi slovami (písaného) korpusu sú v ňom prirodzené, hranica je len jej rozšírením na slovné a neslovné elementy textu (ako interpunkcia; jej prípustnosť vo vstupe je prípustná ako napríklad náznak zmeny intonácie či pauzy v reči, ktorá je pravým zdrojom dát o prvom jazyku pre ľudí). V prípade rečových korpusov platí predchádzajúci odstavec.

Zavedenie hranice do modelu

Explicitné zavedenie hranice (Boundary) tokenov do modelu sa vykoná podobne ako tomu bolo pri hraniciach viet: do grafu sa vloží nový špeciálny uzol, ktorý rozšíri a zjednotí aj uzly begin a end, a cez ktorý bude viesť cesta každej vety (pred a tým aj za) každým (makro)tokenom. Pritom s hranicou sa nepočíta ako s konštituentom viet, slúži len ako oddeľovač slov / zoskupovač znakov v grafe. Stupeň označenia hrán grafu sa zvýši na tri: číslo vety, poradové číslo (makro)tokenu vo vete a poradové číslo znaku v tokene. Hrany vedúce z uzla hranice (pre ktoré je číslo znaku 0, prípadne 1) môžu niesť užitočnejšiu informáciu o dĺžke tokenu, do ktorého vedú.

Afixy ako význačné morfémy

Vyhľadávanie významných vzorov sa priamo opiera o význačnosť hranice: algoritmus si bude „všímať“ podobnosti a opakovania medzi konštituentami v blízkosti hranice. Motiváciou (a inšpiráciou) pre takúto orientáciu bola samozrejme jej priamočiara aplikovateľnosť na slovenčinu (ale nielen ju), kde flexia tvorí snáď hlavnú principiálnu súčasť gramatiky a poradie fráz je voľnejšie.

Na rozdiel od pôvodného prístupu, kedy sa význačné morfémy dôležité pre štruktúru sledovaného jazyka predurčili využitím znalostí experimentátora, je ďalšou úlohou pre algoritmus v súlade s hlavným princípom modelu takéto význačné morfémy hľadať. Keďže hlavnými

slovami (resp. tokenmi aj, alebo aspoň v písanej podobe; obe možnosti sú vcelku prípustné) a kde je v nejakej miere zastúpený flektívny princíp (skloňovania) (asi) v každom jazyku.

V dôsledku toho nemusia byť nasledujúce odstavce a ďalšie zmeny ozaj globálne aplikovateľné a Amigos treba pokladať za konkretizáciu Adiosu (viď pozn. 1).

Na druhej strane aj učiť sa dá učiť, a preto je prijateľným tvrdením, že emergentným správaním sa učiacich sa „algoritmov“ človeka môže byť aj ich vlastná samoorganizácia a konvergencia k výberu len určitých postupov (zameraniu sa len na určitý typ zväzkov v prezentovanom modeli) vedúcich k úspešnému učeniu na prijímaných vnemoch. Takýmto spôsobom by sa mohol mozog zamerať na afixy slov v silne flektívnom jazyku, akým je slovenčina a zavrhnúť vyberanie takýchto kandidátov v iných prípadoch (ide tu o emergentné „zavrhnutie“).

morfémami v slovenčine, ako aj v mnohých iných jazykoch sú afixy, konkrétne predpony a hlavne prípony (hoci sa nájdu aj vpony), zameria sa nový algoritmus na tento druh podobností v jazyku.

Každý z bohatších slovných druhov slovenčiny má priradenú jednu alebo viacero sád prípon (napr. aj sada „e“, „o“ a „šie“ pre príslovky), ktoré sa vzájomne málo prekrývajú a tak sú pomerne jednoznačným identifikátorom gramatickej kategórie (okrem zhody prídavných a podstatných mien; väčšina zámien a číslovky sa dajú považovať za špeciálne prípady podstatných alebo prídavných mien). Neohybné slovné druhy majú menší počet členov a sú používané na význačných pozíciách fráz. Preto možno za konštrukcie vhodné pre slovenčinu považovať postupnosti konštituentov s pevnými morfémami na význačných pozíciách, pričom význačnosť niektorých morfém je bližšie určená na začiatok alebo koniec slova na príslušnej pozícii.

Päť druhov pozícií vo vzore z hľadiska morfológie

Z tohto pohľadu možno rozlišovať päť rozličných reštrikcií na slová na danej pozícii vo vzore:

1. konkrétne slovo. Na pozícii sa musí nachádzať konkrétne pevne zvolené slovo. Použitie v slovenčine je pre predložky, spojky, interpunkciu a podobne. Z pohľadu algoritmu tiež iný vzor.
2. slovo s pevným sufixom. V slovenčine najočakávanejší prípad. Naznačuje použitie slova daného druhu v danom rode, čísle páde atď.
3. slovo s pevným prefixom. V slovenčine zrejme zriedkavé, ale potrebné pre úplnosť a väčšiu všeobecnosť prístupu.
4. oba afixy pevné. Zápor „ne“, tretí stupeň, dokonavosť slovies...
5. ľubovoľné, neurčené. Nevylúčiteľná možnosť, istý fallback do pôvodného správania aktivovateľný pre neflektívne jazyky. V slovenčine je to tiež častá možnosť, napríklad nominatív podstatných mien. Tiež zahŕňa iný vzor gramatiky.

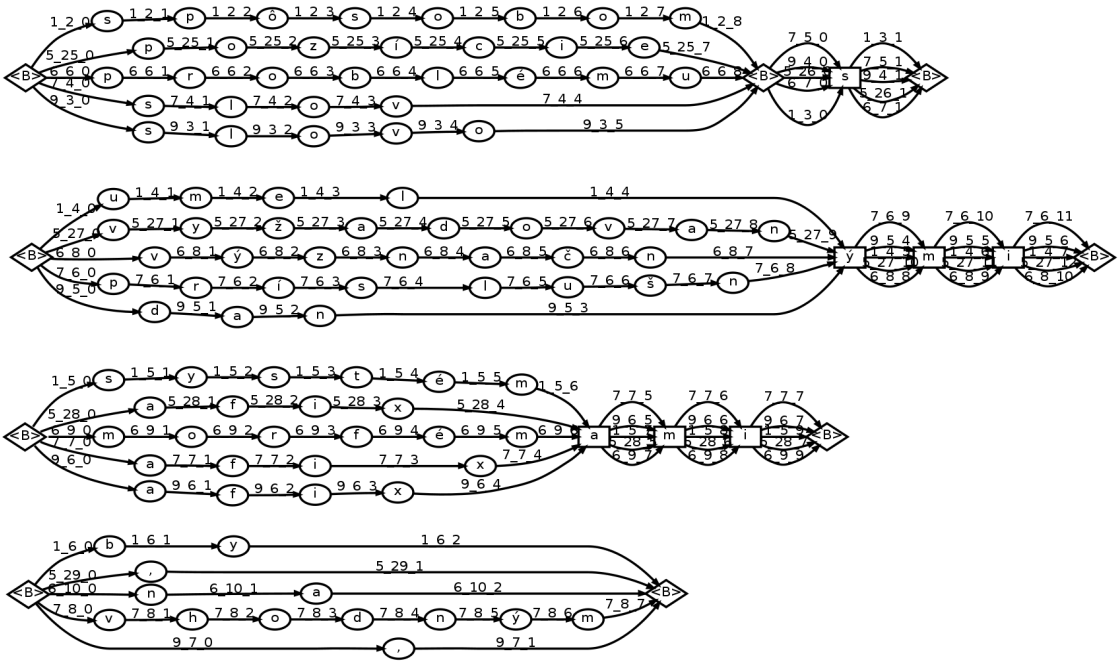
Zamýšľaný nový tvar vzoru

Zamýšľaným tvarom vzorov gramatiky sú teda pôvodné vzory skonkretizované o explicitné hranice (<B>) medzi ich konštituentami rozšírené o niektoré pozície s vyžadovanými afixami, teda skupina prvkov, kde medzera nie je, napríklad „s<B><E1>ými<B><E2>ami“[4], ktorého inštancia je použitá skôr v tejto vete, s príslušným významom konštituentu „majúce objekty jedného z typov z triedy ekvivalencie <E2>, ktoré samy majú vlastnosť opísanú jedným zo slov v <E1>.“

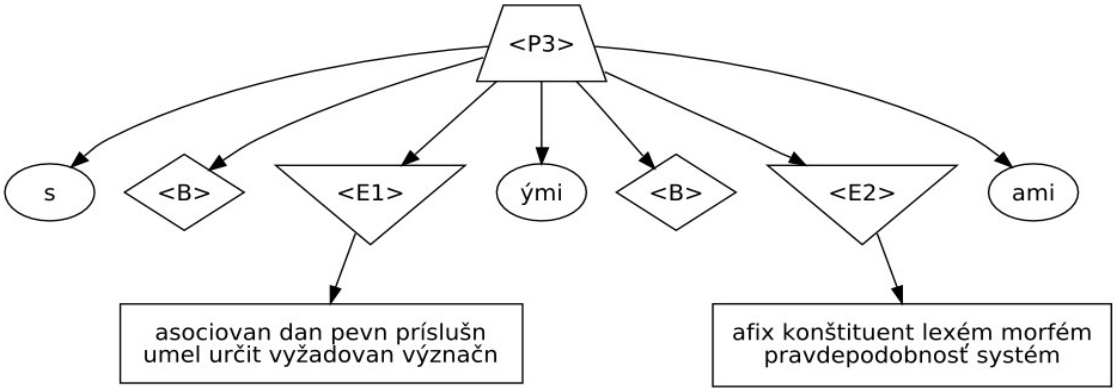
Človeka možno zvádza myšlienka, že by vzory v cieľovej gramatike obdobné príkladu mali najskôr združovať vybrané podstatné mená daného rodu v podľa možnosti všetkých obmenách sufixu (teda aj len s „mi“), podobne aj prídavné mená, ktorých prítomnosť by bola alternatívou v nejakej triede ekvivalencie; potom by sa mohlo na vyššej úrovni rozhodovať medzi rodom a číslom, pričom táto trieda by bola spojená s danou predložkou. Tento pohľad je ale pohľadom konštruktéra jazyka, od prezentovaného modelu ho nemožno a zo strany teoretických východísk ani netreba očakávať.

Cieľom a prezentovaného prístupu nie je len vyriešenie problému s morfémami. Využitím hranice medzi tokenmi odpadajú spomínané starosti so slabičnými vzormi a rôznymi dĺžkami slov. Zároveň je zo zamýšľaného tvaru vzoru vidieť, že vznik podobného vzoru zahŕňa viacero krokov pôvodného modelu. Vzhľadom na požiadavku Adiosu, aby boli prefix aj sufix sekvencie zväzku totožné (resp. totožné s využitím tried) a vzniknúť mohla najviac jedna nová trieda ekvivalencie, umožňuje nový prístup rapídnejší postup, s vyšším stupňom generalizácie a väčšou podporou pre vzory v korpuse.

4 Takéto vzory by mohli nájsť aj svoje praktické využitie v nástrojoch asistujúcich pri korekcii textov v editoroch. V súčasnosti zavedené systémy dokážu detegovať preklepy a časť chýb skloňovania odvolávaním sa na tvary v zabudovanom slovníku. Vhodným rozšírením by bola kontrola zhody napr. v čísle a páde slov (ktorá sa môže stratiť pri revíziách častí viet) potvrdzovaná prítomnosťou takého vzoru v gramatike a jeho očakávateľnosti na danom mieste vo vete.



Obrázok 2: Príklad časti signifikantného zväzku ciest RDS grafu v modeli Amigos. Cesty grafu boli rozvinuté do osobitných uzlov, v skutočnosti je v RDS grafe jediný uzol pre každý znak. Medzi každým tokenom sa cesta vráti do uzla hranice (<B>). Signifikantný vzor v tomto zväzku tvoria podcesty od druhej inštancie hranice po piatu. Vzorom vyžadované prvky sú označené obdĺžnikovými uzlami Vzor je zobrazený na obr. 3. Vyhľadávaním podobných zväzkov sa nachádzajú významné syntaktické pravidlá (zhodný prívlastok, väzba predložky s pádom), pričom sa obchádza problém rôznych dĺžok slov a vysokej variability okolitých fráz (prvé a posledné slovo zobrazených ciest).



Obrázok 3: Zamýšľaný signifikantný vzor. Ďalšie prvky tried ekvivalencie pochádzajú z ostatných ciest signifikantného zväzku z obr.2.

Opodstatnené modifikácie

Možné rozšírenia navrhovaného postupu zahŕňajú na jednej strane možnosť viacerých úrovní hľadania vhodného rozdelenia tokenov, resp. pospájania znakov medzi dvoma nasledujúcimi prechodmi cez uzol hranice vo vetách, na strane druhej (ktorá môže podporiť prvú a naopak) sa afixy víťazného zväzku dajú týmto víťazstvom licencované za pravdepodobne pravé význačné afixy jazyka, ktoré bývajú v rámci gramatiky globálnejšie použiteľné a po ich nájdení je možné nahradiť všetky ich príslušné výskyty v korpuse novým uzlom.. Prvá strana umožňuje rozšírenie prezentovaného prístupu na viacero stupňov afixov, čo je potrebou pri aglutinačných jazykoch (ktoré ku kmeňu bežne prilepujú mnoho stupňov afixov, čím vznikajú až tisíce tvarov daného slova; príkladom takého jazyka je fínčina), alebo na kompozitá (veľmi využívané v nemčine). Oba druhy použití sa nájdu aj v slovenčine. Druhá strana zase umožňuje rýchlu redukciu problému s význačnými morfémami na iniciálne iterácie spracovania príslušných viet, odkiaľ by sa pokračovalo podľa pôvodného algoritmu. Nový uzol sa dá jednak považovať za token pre Adios, dvak ide o validný vzor a aj konštrukciu s priraditeľným čiastkovým významom.

Principiálnym rozhodnutím učineným autorom tejto práce a možno z rovnakých dôvodov aj autorov Adiosu je nezaoberať sa tvorbou rekurzívnych vzorov v gramatike. Rekurzia, ako ju vidieť v konštruovaných gramatikách prirodzených jazykov, je (ideálnym) logickým konštruktom zjednocujúcim viaceré možné úrovne rozvitia frázy rovnakým spôsobom, čím napomáha jej väčšej prehľadnosti a manipulovateľnosti. Hoci si ľudia dokážu predstaviť potenciálne neohraničený počet použití rekurzie vo vetách, nezvyknú pri nej ísť nad rozumne spracovateľnú (malú) hranicu. Gramatika má slúžiť na podchytenie skutočného jazyka (množiny gramatických viet), preto je využitie rekurzie v realistických gramatikách nadbytočné.

Hoci je rekurzia v gramatike na obtiaž, nie je úprava výsledného tvaru (či revízie už počas získavania) gramatiky porovnávaním a zjednocovaním, či naopak rozdeľovaním vzorov, resp. tried ekvivalencie neočakávaným javom pri učení sa jazyka. Súčasné modely sú ale postavené na jednosmernom greedy redukovaní dát pomerne výrazne vyžadujúcom stratu informácie o presnom priebehu subsumovaných ciest (aby neboli subsumované opakovane).

Vyhľadávanie kandidátskych zväzkov pomocou význačnej cesty

Kostra algoritmu akvizície vzorov sa od pôvodného modelu samozrejme principiálne nelíši. Zvolený tvar hľadaných vzorov ale vyžaduje niektoré ďalšie konkretizácie a modifikácie algoritmu, špeciálne spôsob vyberania kandidátskych zväzkov a rozhodovanie o signifikantnosti vzorov.

V každej iterácii algoritmu sa zväzky hľadajú podľa jedinej vytýčenej podcesty v grafe, Vyhľadávacej Cesty (VC). Dĺžka tejto podcesty môže byť z praktických dôvodov ohraničená, ale môže ňou byť aj aktuálna reprezentácia niektorej celej vety korpusu. Takto sa môže postupne vyskúšať každá veta korpusu dookola, alebo je veta určená náhodným výberom spomedzi všetkých viet. Tým sa dá vyhnúť závislosti výsledku na poradí viet vo vstupe.

Odôvodnením tohto postupu je predpoklad, že každá veta korpusu vznikla inštancovaním vzorov cieľovej gramatiky, preto je rozumné ich hľadať hoci aj v nej. Zároveň by mali byť jednotlivé inštancie vzoru medzi sebou rovnocenné a hľadanie ciest podobných s vyhľadávacou ich veľkú časť vyhľadať. Zároveň sú viaceré navzájom si podobné vzory v gramatike očakávaným a psycholingvisticky prijateľným stavom. Do tretice má každý signifikantný vzor právo vzniku, a tak nie je nutnosťou vyhľadať v každej iterácii aktuálne globálne najsignifikantnejší vzor celého (obrovského v porovnaní s očakávaným rozšírením priemerného vzoru) korpusu. Navyše je väčšina vzorov vďaka ich závislosti na kontexte relatívne nezávislá, čím by sa zbytočne vytváralo veľa kandidátov vo viacerých iteráciách prakticky bezozmeny.

Identifikácia kandidátov na význačné afixy

V každom slove vyhľadávacej cesty sa algoritmus najprv pokúsi nájsť afixy. Preto postupne rozdelí dané slovo na rôzne zreťazenia prefixu, kmeňa a sufixu, kde len kmeň je povinnou súčasťou slova. Či je rozdelenie vhodné sa zistí prehľadaním príslušnej časti grafu korpusu (tvorenej cestami medzi znakmi slova). Myšlienkou je, že dobré oddelenie afixov od kmeňa má za následok kompaktné reťazce s dobrou podporou v dátach. Algoritmus teda medzi sebou porovná skóre pre rôzne hypotézy o rozdelení slova. Vierohodnosť danej hypotézy sa odhadne podľa relatívnych početností daného prefixu (len na začiatku slova), kmeňa (všetky výskyty) a sufixu (na konci ľubovoľného slova). Ďalej sa ale dostane len také rozdelenie, ktoré prekonáva globálnu hranicu určenú parametrom algoritmu. Obdobné správanie by sa dalo

pravdepodobne očakávať aj od pôvodného modelu, ak by mal za úlohu vytvoriť gramatiku z korpusu písmen oddelených medzerami.

Zo syntaktického hľadiska sa považujú slová daného ohybného slovného druhu, resp. ich tvary so zhodnými gramatickými kategóriami za rovnocenné. Preto sú všetky výskyty slov s príslušnými afixami vhodným základným výberom z grafu/korpusu, na ktorom možno hľadať nový signifikantný vzor.

Vyššie bolo uvedených päť kategórií reštrikcií na prvky očakávateľných na danom mieste vzoru predstavovaného modelu. Aby sa umožnila rovnaká štartovacia pozícia pre každú z týchto kategórií a tak obmedzila možnosť kategorickej chyby v rozdelení slova, vytvorí sa postupne kandidátsky zväzok s predpokladom každej kombinácie výberu z týchto reštrikcií. Preto sa pre každú pozíciu zistia štyri rôzne základné výbery: všetky slová s totožným najlepším prefixom, sufixom a najlepšej dvojice afixov slova a navyše všetky výskyty celého slova (tvaru) z vyhľadávacej cesty. Uvedené pravidlo sa pravdaže nevzťahuje na vzory či tie konštituenty vyhľadávacej cesty, ktoré sú príliš krátke.

Tvorba kandidátskych zväzkov

Pre každú podcestu vyhľadávacej cesty a každú prípustnú kombináciu reštrikcií sa nájde kandidátsky zväzok. Je to množina všetkých ciest rovnakej dĺžky ako skúmaná podcesta, ktorých jednotlivé konštituenty patria do základného výberu príslušného konštituentu tejto podcesty pri danej reštrikcii. Ak sa daný zväzok stane víťazom, považuje sa to za „dôkaz“ o tom, že dané rozdelenie slova na afixy je na danom zväzku správne a požadované.

Čo sa týka piatej kategórie, teda ľubovoľného slova, tam zrejme základný výber nie je potrebné vytvárať, tvoria ho všetky (makro)tokeny korpusu. Zo zjavných dôvodov zabránenia prílišnému zovšeobecneniu a chybám je potrebné obmedziť použitie tejto reštrikcie na maximálne jednu v každom zväzku, pričom táto nie je vhodná na okraji, kde by dochádzalo k ťažko ovládateľnému pripájaniu slov vhodnejších v potenciálne viacerých susedných vzorov.

Signifikantnosť kandidátskeho zväzku

Vyhodnocovanie signifikantnosti daného zväzku je porovnateľné s postupom Adiosu, až na jedno zjednodušenie. Toto sa vyhýba nutnosti vyhľadať v rámci jednej iterácie všetky výskyty všetkých slov a bigramov všetkých ciest zo zväzkov (potenciálne ide o všetky slová

zo základných výberov), keďže pracovať s cestami v zväzku ako postupnosťami znakov s rôznymi dĺžkami je nerozumné, čo by bolo pamäťovo a časovo náročné (ešte raz, korpus je rozdelený po znakoch) a tak obmedzujúce na veľkosť korpusu. Preto sa podobne ako pri hľadaní signifikantného vzoru v signifikantnej ceste modelu ADIOS2 berú základné výbery ako jeden konštituent, ktorého výskyt je zjednotením výskytov prvkov z neho. Vzorec signifikancie zväzku sa tak premení na:

$$S(Z) = P^{(k)}(Z) \log \frac{P^{(k)}(Z)}{P^{(2)}(Z)}$$

$$P^{(k)}(Z) = P(v_1)\,P(v_2|v_1)\,P(v_3|v_1 \to v_2) \cdot \ldots \cdot P(v_k|v_1 \to v_2 \to \ldots \to v_{k-1})$$

$$P^{(2)}(Z) = P(v_1)\,P(v_2|v_1)\,P(v_3|v_2) \cdot \ldots \cdot P(v_k|v_{k-1})$$

$$kde\ Z = v_1 \to v_2 \to \ldots \to v_k = \{c_1 \to c_2 \to \ldots \to c_k \mid \forall i : c_i \in v_i\}$$

Implementačný pohľad na algoritmus

Z pohľadu implementácie nie je algoritmus principiálne zložitý. Základom implementácie grafovej štruktúry sú uzly, jednoznačne identifikovateľné a vyhľadateľné podľa svojho čísla, z ktorých každý obsahuje pole čísel hrán, ktoré z neho vychádzajú a uzlov, do ktorých tieto hrany vedú. Kladné číslo hrany je uložené v jedinej premennej s troma bitovými komponentami (postupne pre číslo vety, tokenu a znaku v tokene). Hrany uzla sú usporiadané podľa ich čísel, teda poradia výskytu daného znaku v korpuse, umožňujúc tak binárne vyhľadávanie. K vetám a (makro)tokenom korpusu sa pristupuje z uzla hranice, kde sú tieto tiež uvedené v súvislom poradí.

Základnou operáciou využívanou pri hľadaní príslušných ciest v grafe je prienik dvoch utriedených postupností do novej utriedenej postupnosti s prípadným konštantným rozdielom medzi príslušnými hodnotami (napr. jedného znaku alebo slova). Základnými takýmito postupnosťami sú bigramy znakov, ktoré vzniknú prefiltrovaním poľa hrán vedúcich z prvého konštituentu. Tieto sa v rámci jednej iterácie môžu vytvárať podľa potreby a kešovať, lebo udržiavať všetky bigramy grafu oddelene zbytočne komplikuje implementáciu. Zoznam zložitejších (dlhších) ciest danej dĺžky obsahuje význačné (napr. prvé) hrany v nich. Takto sa dajú technikami dynamického programovania vhodne usporiadať jednotlivé operácie (napríklad vyhodnocovať postupne sa predlžujúce kandidátske zväzky), čím sa odstráni nutnosť viacnásobného počítania tých istých dát. Sledované štatistiky sú jednoducho dĺžky jednotlivých zoznamov.

Vzhľadom na to, že sa vnútorná štruktúra vzorov ani obsahy tried ekvivalencie v modeli nevyužívajú, je možné tieto dáta odpútať od RDS grafu a reprezentovať ako textové reťazce, podobne ako v príklade zamýšľaného vzoru. Dôležitým bodom správneho presmerovávania ciest grafu je najskôr overiť, že žiadna časť danej inštancie vzoru už nebola subsumovaná tým istým vzorom skôr. Súčasťou presmerovávania je mazanie hrán v rámci a znižovanie hodnôt (čísla slova) vo všetkých nasledujúcich hranách za výskytom vzoru. Z dôvodu zachovania invariantov (nutných pre binárne vyhľadávane) je nutné uskutočňovať dané operácie vo fázach. Hrany na vymazanie sa preto najskôr označia (napr. nastavením odkazu nasledovného uzlu na null), a až nakoniec skompaktovať.

Hľadanie zhody medzi frázami a vzormi

Záver tejto kapitoly bude venovaný podobne ako v kapitole 3.2 parsovaniu viet pomocou vzorov gramatiky. Tu sa tiež naskytuje možnosť úprav interpretácie aktivácií a spôsobu ich získavania vyplývajúca zo zamýšľaného tvaru vzorov nového modelu.

Prítomnosť nejakého slovného konštituentu v sekvencii vzoru (naproti jeho prítomnosti v triede ekvivalencie) je odrazom silnej viazanosti vzoru na jeho prítomnosť vo fráze, ktorá je jeho inštanciou. Preto neprítomnosť takéhoto konštituentu, teda napr. sufixu alebo predložky svedčí o kategorickej nepoužitelnosti vzoru na danú časť parsovanej vety. Naproti tomu neprítomnosť (kmeňa) slova v príslušnej triede ekvivalencie nie je dôvodom odmietnuť daný vzor, ak je prienik s vyžadovanými konštituentami dostatočný.

Proces získavania aktivácií môže aj vďaka (jednosmernej) hierarchii vzorov prebiehať postupným hľadaním zhody so vzorom v každom mieste analyzovanej vety v poradí získania týchto vzorov. Vychádzajúc z danej pozície vo vete, porovná sa príslušná časť vety s predpisom na každom mieste vzoru.

Ak je predpísané priamo nejaké slovo, môže sa pokračovať porovnávanie s ďalším slovom vety len ak je toto slovo aj na danom mieste vety (s hodnotou zhody jedna). Ak vzor predpisuje slovo s danými afixami, pokračuje sa len ak sa v slove nachádzajú, pričom hodnota zhody (daná časť jednotky) závisí od prítomnosti kmeňa slova v danej triede ekvivalencie.

Predpísaný výskyt iného vzoru znamená preskúmanie všetkých výskytov tohto vzoru na danej pozícii, s mierou zhody prevzatou z daného výskytu, pričom sa pokračuje v porovnávaní slovom nasledujúcim za týmto výskytom. Trieda ekvivalencie na danom mieste vzoru sa ošetruje ako zhoda postupne s každým jej prvkom.

Takýmto spôsobom sa zistia všetky čiastočné výskyty daného vzoru (využiteľné v jemu nadradených vzoroch). Mierou zhody výskytu vzoru je potom súčet všetkých čiastkových zhôd pre jednotlivé časti sekvencie vzoru, a tak zhruba zodpovedá dĺžke frázy, na ktorej sa tento výskyt nachádza (v prípade úplnej zhody je jej rovná). Víťazným vzorom resp. jeho výskytom (z viacerých možných) je ten s maximálnou mierou zhody. Mierou úspešnosti gramatiky postihnúť analyzovanú vetu sa potom stane pomer tejto hodnoty s dĺžkou vety.

5 Experimenty a ich vyhodnotenie

5.1 Učenie sa vzorov náhodných gramatík

Na preverenie schopnosti prezentovaného modelu získavať vzory obsiahnuté vo vstupnom korpuse sa vykonalo niekoľko pokusov na malých a zjednodušených, a tým aj lepšie vyhodnotiteľných umelých príkladoch. Z dôvodu širšieho záberu do priestoru možných kombinácií konštrukcií sa testovacie gramatiky konštruovali náhodne. Aby sa mohla vyhodnotiť schopnosť gramatiky rozpoznávať afixy, boli tieto gramatiky modelované podľa zamýšľaného tvaru získaných gramatík.

Generovanie náhodných gramatík

Za týmto účelom sa najskôr vygeneroval lexikón zdieľaný všetkými gramatikami. Lexikón bol tvorený štyrmi triedami konštituentov: prefixy, sufixy, pomocné slová a napokon všeobecné slová a kmene. Použité bolo náhodné rozdelenie dĺžok slov a zastúpenia písmen v slovách. Tieto kategórie sa rozlišovali kvôli rozdielnym hraniciam dĺžky konštituentu, zhrnutom v tab. 1.

Trieda	***Počet***	***Minimum***	***Maximum***
Prefix	100	2	4
Sufix	100	2	4
Pomocné slovo	100	1	6
Kmeň slova	1000	2	8

Tabuľka 1: Mohutnosť a hranice dĺžky tried v lexikóne

Vzory gramatík sa vyberali s dĺžkami od 2 do 5 slov a veľkosťami tried ekvivalencie od 2 do 10 prvkov. Pre každú pozíciu sa zvolila jedna z piatich tried z kapitoly 4.2 a podľa nej sa zvolili jednotlivé časti z príslušných tried lexikónu. Do kategórie „celé slovo“ sa vyberali pomocné slová a s pravdepodobnosťou 25% aj predchádzajúce vzory. Rovnako tvorili iné vzory maximálne 25% prvkov tried ekvivalencie. Použité rozdelenia sú zhrnuté v tab. 2. Takýmto spôsobom sa vytvorilo 100 umelých gramatík po 30 vzorov.

Parameter	*Hodnota alebo rozdelenie*
Dĺžka vzoru	*2*: 40 %; *3*: 45 %; *4*: 10 % *5*: 5 %
Veľkosť triedy ekvivalencie	V pomere x + 3 pre 1 < x < 7 a 12 – x pre 7 <= x <= 10
Kategória pozície vo vzore	VT : PF : SF : AF : AW = 2 : 1 : 2 : 1 : 1
Podiel vzorov	25% vo VT a max. 25% v AW
Počet vzorov	30
Počet gramatík	100

Tabuľka 2: Parametre generovania gramatík (VT – verbatim = celé slovo / iný vzor; PF, SF, AF = rovnaký prefix, sufix, resp. oba afixy s triedou ekvivalencie kmeňov; AW = ľubovoľné slovo či vzor ako konečná trieda ekvivalencie)

Kategória	*Max. hĺbka*	*Počet pravidiel*	*Počet koreňov*	*Expresivita*
Minimum	2	73	16	934
Maximum	5	105	28	4 779 572 228
Medián	3	89	23	39 469
Priemer	3,11	89,66	23,19	58 240 884

Tabuľka 3: Vybrané (nezávislé) vlastnosti generovaných gramatík. Pravidlá sú spolu pre vzory aj triedy ekvivalencie. Expresivita je počet viet generovateľných gramatikou (z koreňov; rôznym výberom v triedach ekvivalencie). Ako vidno, skúmané gramatiky nereprezentujú celkom triviálne jazyky.

Testovanie algoritmu

Z každej takto generovanej gramatiky sa vyprodukoval korpus 10 000 viet (resp. fráz), na ktorom sa pustila implementácia modelu. Táto implementácia vyhľadávala afixy dĺžok od 2 do 5, pričom ponechávala na kmeň slova 2 znaky. Ako globálna minimálna hranica akceptovateľnej signifikancie vzoru bola použitá hodnota 0,0, ktorá zodpovedá rovnosti medzi vierohodnosťami nulovej a alternatívnej hypotézy. Dôvodom tejto voľby bol cieľ preskúmať všetky potenciálne získateľné vzory a apriórna neurčiteľnosť „rozumnej“ hodnoty. Algoritmus zastavil po 1 000 neúspešných pokusoch vytvoriť nový vzor podľa náhodne vybranej vety.

Získané gramatiky sa otestovali pomocou miery zhody s príslušnými trénovacími korpusmi (Zhoda $G_{learned}$ a L_{train}) a naopak zhody korpusu 10 000 náhodne vybraných viet generovaných týmito gramatikami a pôvodnou gramatikou (G_{train} a $L_{Learned}$. Miera zhody sa merala dvoma spôsobmi: v kategorickom móde sa vyžadovala úplná zhoda celej vety s nejakým vzorom gramatiky (teda že daný vzor vetu generuje). V druhom sa spočítali podiely víťazných aktivácii vzorov a počtu slov v analyzovanej vete. Zhrnutie výsledkov tohto experimentu je uvedené v tab. 4.

Kategória	$G_{learned}$ *a* L_{train}		G_{train} *a* $L_{Learned}$	
	Presne	***Čiastočne***	***Presne***	***Čiastočne***
Minimum	3860	6008,01	2266	3488,58
Maximum	8288	9342,83	9688	9730,60
Medián	6334	7957,11	7328	7768,04
Priemer	6328,53	7884,53	7232,32	7566,85

Tabuľka 4: Zhrnutie výsledkov úspešnosti akvizície umelých gramatík

Úspešnosť algoritmu

Dosahované hodnoty úspešnosti neboli závislé na expresivite ani na maximálnej úrovni hierarchie vzorov v trénovacej gramatike. Obe minimá (čiastočne a presne) sa dosahovali v tých istých gramatikách, maximá boli v rôznych. Takisto úspešnosť v jednom smere testu nebola v priamej zhode s úspešnosťou v druhom.

Vlastnosťou naučených gramatík je znížená maximálna výška vzoru (na max. 2) a väčší, asi dvojnásobný počet pravidiel gramatiky. Expresivita naučených gramatík zodpovedá veľkosti trénovacích korpusov a nie je jednoznačne závislá od expresivity trénovacej gramatiky (hoci je vždy menšia). Expresivita nemá degradačný účinok na úspešnosť (napr. najbohatšia gramatika mala 76% a 90% úspešnosť). Tieto vlastnosti sú zhrnuté v tab. 5.

Kategória	*Počet vzorov*	*Počet pravidiel*	*Nárast počtu pravidiel*	*Expresivita*	*Pomer expresivít v %*
Minimum	39	112	1,21	492	< 0,01
Maximum	169	368	3,72	133 085	94,53
Medián	77	179	1,98	3 637	12,02
Priemer	82,24	184,92	2,06	9 300,34	22,88

Tabuľka 5: Vlastnosti naučených gramatík v porovnaní s ich trénovacími (tab. 3)

Závislosť úspešnosti od veľkosti vstupu

Uvedené výsledky sú uspokojivé, ale nie optimálne. Preto sa na jednej z menej úspešných (číslo 85; úspešnosť v poradí z tab. 4 4663; 7294,12; 5635 a 6316,33) s vhodnou expresivitiou (5 488; max. úroveň 3). Na tejto gramatike sa sledovala závislosť úspešnosti v teste od veľkosti vstupného korpusu. Ďalej sa sledovala možnosť generovať trénovacie (a následne aj testovacie) korpusy z každého vzoru gramatiky, nielen z tých, ktoré sú v koreňoch and–or stromov v grafovej reprezentácii gramatiky.

Veľkosť vstupu (× 1000)	*Expre-sivita*	*Len z koreňov*		*Z ľubovoľného vzoru*	
		$G_{learned}$ a L_{train}	G_{train} a $L_{Learned}$	$G_{learned}$ a L_{train}	G_{train} a $L_{Learned}$
2	1380	6470	4748	6032	4983
5	1495	5915	4706	5900	5638
10	1723	5558	4679	6510	5151
15	1839	5253	4706	6542	5449
20	1675	5481	4206	6315	5099
30	1668	5741	4765	6253	5108
40	1701	5492	4601	5295	4988
50	1527	5156	4713	6415	5086
75	1683	5888	4580	6381	5352
100	1697	5887	4592	6255	5149
Priemer	1639,8	5684,1	4629,6	6289,8	5200,3

Tabuľka 6: Testovanie rôznych veľkostí korpusu na gramatike 85. Uvedené hodnoty sú presnej zhody. Hodnoty čiastočnej zhody boli tiež takmer nemenné, priemerne 6604, 7414, 6722 a 7452.

Z výsledkov v tab. 6 vidieť, že úspešnosť gramatiky sa s veľkosťou vstupného korpusu nemení (čo na druhej strane znamená, že sa vzor môže vytvoriť aj keď je zahalený inštanciami iných) a generovanie zo všetkých vzorov akvizícii mierne napomáha.

Identifikácia význačných afixov

V rámci predchádzajúceho experimentu sa sledovalo aj pokrytie vyžadovaných konštituentov trénovacích vzorov v získaných. Ani tu sa nedosahovali rozdiely v závislosti od veľkosti vstupu. Pôvodná gramatika definovala 19 prefixov, 20 sufixov a 14 význačných slov. Získané gramatiky definovali v prieniku 10, 14 a 33 prvkov každej kategórie, z toho 8, 11 a 6 vyskytujúcich sa v pôvodnej gramatike. Individuálne boli početnosti prienikov s pôvodnými 12, 13 a 6 prvkov (s minimálnou variáciou o jednotku). Tento počet sa zvýšil, napr. na 16, ak sa rátal aj kratší získaný sufix (aj získané sufixy boli dĺžky 4). Tieto hodnoty sa mierne pohli o 1–2 rôznymi smermi, ak sa trénovacie vety generovali zo všetkých vzorov, pričom identity (aj správne) identifikovaných afixov a slov sa inak nezmenili (buď nejaké pribudli, alebo ubudli).

Druhá generácia učenia

Tieto aj predchádzajúce výsledky naznačovali, že získané vzory a jazyky predstavujú tú časť náhodných gramatík, ktorú sa je algoritmus schopný naučiť. Toto nepredstavuje z teoretického hľadiska nepríjemnosti: jazyky sa vyvinuli tak, že sa *dali* naučiť a zároveň prijateľný všeobecný algoritmus učenia nemusí dokázať úspešne pracovať na ľubovoľnom vstupe. Na podporu tejto hypotézy sa na korpusoch generovaných získanými gramatikami natrénovali nové gramatiky a porovnali navzájom. Výsledky zhrnuté v tab. 7 a 8 sú jednoznačnou podporou tejto hypotézy: úspešnosť novej sady gramatík na pôvodných dátach príliš neutrpela, zhoda s prvou generáciou je ale omnoho lepšia.

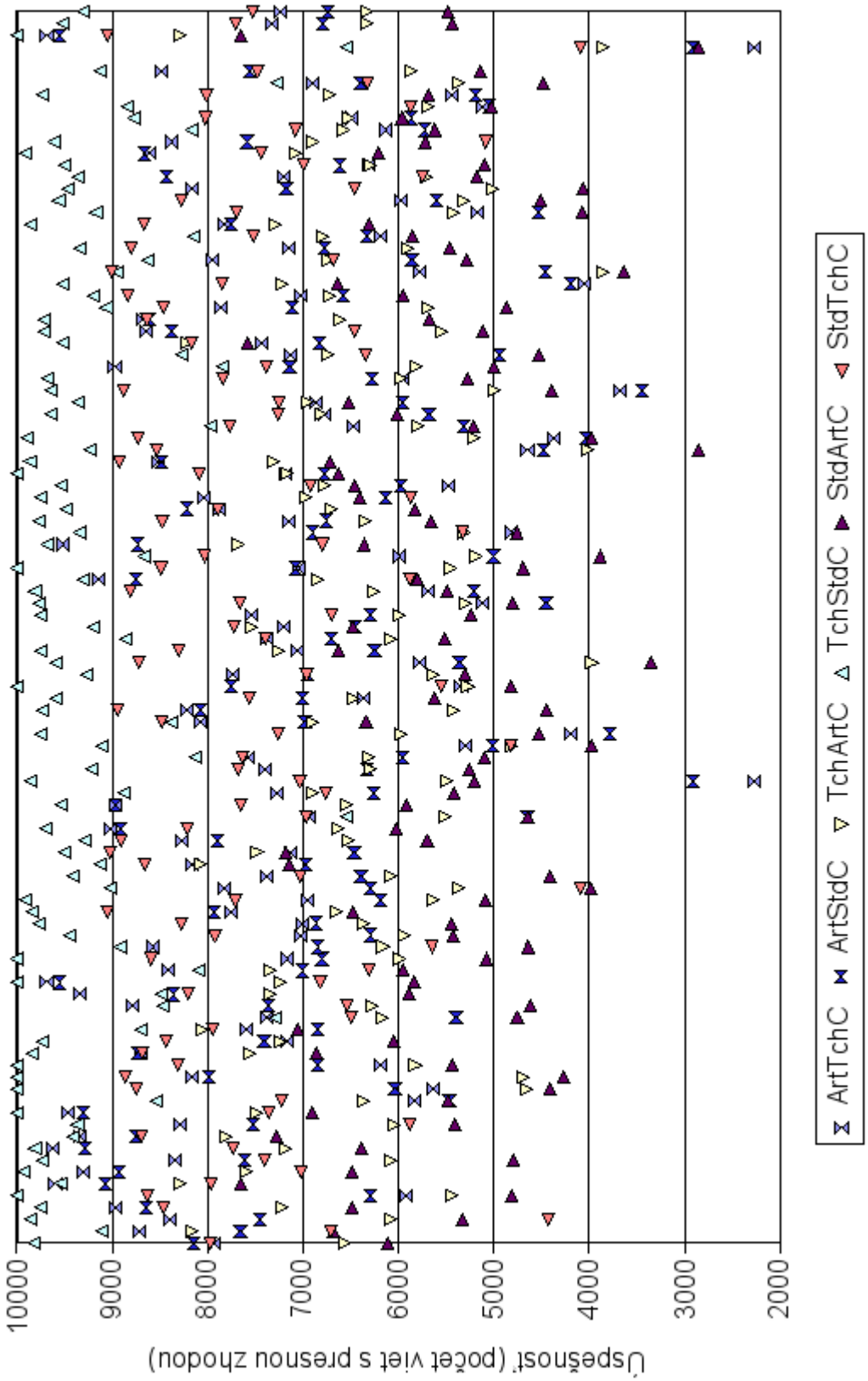
Kategória	$G_{learned}$ ***a*** L_{train}		G_{train} ***a*** $L_{Learned}$	
	Presne	***Čiastočne***	***Presne***	***Čiastočne***
Minimum	2856	5760,59	2919	3672,33
Maximum	7652	8671,33	9560	9560,00
Medián	5440	7104,12	6788	7177,09
Priemer	5478,44	7147,56	6744,32	7143,49

Tabuľka 7: Zhoda medzi umelými gramatikami a druhou generáciou získaných

Kategória	$G_{learned}$ ***a*** L_{train}		G_{train} ***a*** $L_{Learned}$	
	Presne	***Čiastočne***	***Presne***	***Čiastočne***
Minimum	6546	6642,75	4096	4443,67
Maximum	10000	10000,00	9056	9398,10
Medián	9520	9588,63	7706	8305,50
Priemer	9312,42	9387,51	7530,25	8053,10

Tabuľka 8: Zhoda medzi generáciami získaných gramatík

Všetky dáta o presnej zhode (dáta o čiastočnej zhode majú podobný charakter) sú vizualizované v grafe 1, kde je zreteľná hlavne spomínaná nezávislosť úspešnosti na expresivite danej umelej gramatiky.



Graf 1: Úspešnosť (presne) 100 gramatík (prvý komponent názvu) na korpuse 10 000 viet inej gramatiky (druhý komponent). Art sú umelé gramatiky, Tch a Std sú prvá a druhá generácia natrénovaných. Dáta sú usporiadané podľa expresivity (počtu generovateľných viet) príslušnej umelej gramatiky. Zjavná chaotickosť hodnôt a väčšia zhoda medzi získanými gramatikami značí závislosť na zložitosti konštrukcií a nie ich veľkosti.

5.2 Akvizícia slovenského jazyka

Ako hlavný experiment práce algoritmu sa vyskúšalo získavanie vzorov na reálnom korpuse slovenského jazyka. Použitá bola malá časť Slovenského národného korpusu. Tento korpus je nešpecializujúcou sa zbierkou textov súčasného slovenského jazyka z rôznych zdrojov a rôznych tematických oblastí, štýlov a žánrov. Použitím takéhoto korpusu sa tak obtiažnosť úlohy pre model zvýšila tretím spôsobom: po strate informácie o význačných morfémach v jazyku algoritmus musí získavať vzory z (obtiažnejšieho) slovenského korpusu, ktorý navyše neobmedzuje cieľovú skupinu a tým aj zložitosť použitých viet (využitie korpusu Childes v experimentoch s Adiosom týmto nepodlieha kritike, takéto vety boli vhodné pre možnosť vzťahovania paralel s reálnou akvizíciou prvého jazyka u detí).

Použitý korpus tvorilo 255 986 automaticky rozdeľovaných viet z viacerých (z hľadiska metadát korpusu bližšie nešpecifikovateľných) dokumentov, ktoré spolu mali 3 730 279 tokenov (slová, symboly, interpunkcia) a 15 963 552 znakov bez medzier. Na tomto korpuse bol natrénovaný algoritmus za rovnakých podmienok ako v predchádzajúcich pokusoch. Program zastavil po asi deviatich dňoch nepretržitej činnosti, keď korpus zredukoval na 1 967 805 tokenov aplikovaním .156 495 vzorov.

Signifikantnosť všetkých vytvorených vzorov sa nezávisle od pociťovanej relevantnosti a prijateľnosti pohybovala rádovo medzi 10^{-3} a 10^{-6}. Za signifikantné sa na jednej strane považovali pravidelné spojenia častých slov, na strane druhej (väčšina druhej polovice získaných vzorov) sa vyberali dlhšie postupnosti jedinečných slov (ktoré boli týmto činom signifikantne previazané).

Algoritmus výrazne preferoval možnosť zaradiť ľubovoľné slovo medzi dvojicu viazanejších pozícií, čím sa dosiahla väčšia bohatosť zväzku za cenu jeho prílišného zovšeobecnenia; počet prvkov príslušnej triedy ekvivalencie ale bol menší ako variabilita viazaných. Aj vďaka (vedomej) vysokej bohatosti vzorov na najnižších úrovniach (postačovala zhoda afixov, bez potreby vonkajšieho kontextu) sa generalizačná schopnosť vzorov vyššie v hierarchii prudko zvyšovala, čím utrpela schopnosť týchto vzorov docieliť zhodu na vzdialenejších pozíciách v týchto vzoroch.

Medzi významné pozorované triedy správania (získané vzory) algoritmu patria:

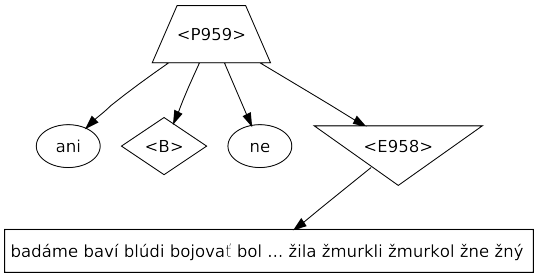
- Slovné spojenia (nie nutne idiómy) okrem základných ako použitie pomocného slovesa byť, zvratných sa a si, častých spojení častíc a zamien sa našli *dobre teda, škoda reči, celý deň, školský rok, pravý čas, modré oči, svetlé vlasy, modrý kameň, NR SR, jedného dňa, prosím vás*, alebo aj *mnohé iné*
- Zhody v pádoch medzi prídavnými a podstatnými menami, medzi osobou a slovesom či pomocným slovesom byť a časom, podmienené frázy a podobne (viď obrázky). Získať takéto vzory bolo hlavným cieľom experimentu, keďže samostatné vyhľadávanie afixov a ich zhôd (na úkor závislosti vzorov na kontexte) bolo hlavným motívom modifikácie pôvodného modelu.
- Ako test bola algoritmu poskytnutá možnosť vyberať prefixy slov ako vyžadované súčasti vzorov. Okrem vhodných vzorov ako „*nikdy ne{bol, …, žila}*“ alebo „*naj{lep,hor}šie bude*“, či predpôn slovies sa prefixy vyberali veľakrát bez výrazného súvisu medzi jednotlivými frázami v príslušnom zväzku, alebo sa ako prefix vybral kmeň slova („*všetk*“). Podobnosti medzi prefixmi sú v jazyku (a vstupnom korpuse) početné, preto algoritmus nemal dôvod ich odmietať.
- vo vstupnom korpuse sa nachádzali aj anglické vety, na ktorých možno badať podobnosť práce nového algoritmu na jednoduchšom jazyku s Adiosom. Medzi vybrané slovné spojenia a vzory patrili: *thank you, custard pie, greatest hits, play list, let it, jump back, yoko ono, Santa Cruz, blind eye, talking back, for you, want to, wishbone ash, hang fire, passion play, higher love*, zo vzorov s triedami ekvivalencie to bolo napríklad „*too {young, old, try} to*“ (kde try bolo pridané kvôli chybe, keď malo byť to try to) a.i.

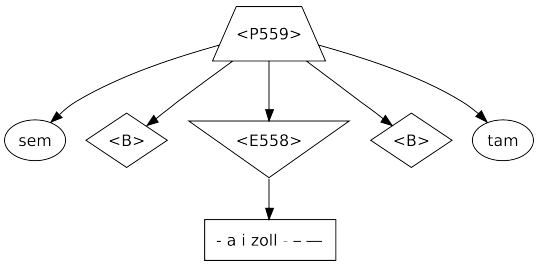
Komentár dosiahnutých výsledkov

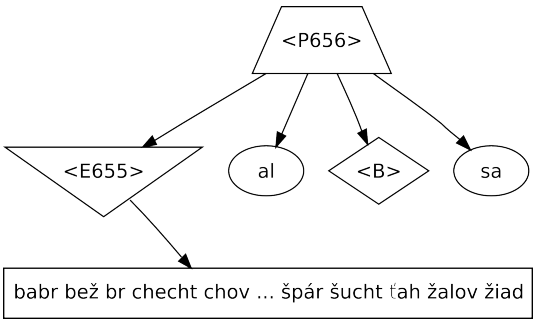
Závery, ktoré vyplývajú z uskutočneného experimentu hovoria, že predchádzajúca informácia o význačných morfémach v modeli Adios nie je potrebná, ale zároveň nie je ani potrebné vyhnúť sa jej obtiažnym spracovávaním korpusu až po znakoch, keďže sa tieto morfémy (a potvrdenie o ich význačnosti v podobe zhody s inými význačnými morfémami) dajú rozpoznať vytváraním kandidátskych zväzkov vyžadujúcich zhodu v okolí medzier, teda na vzdialenejších (a ľubovoľne vzdialených) pozíciách v postupnostiach znakov.

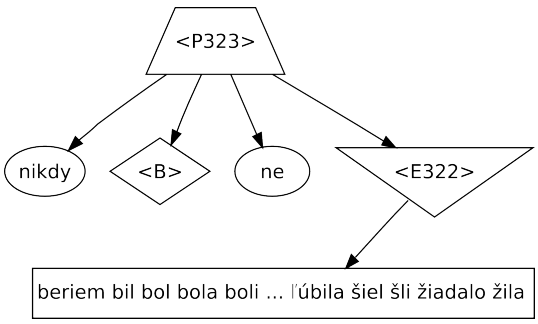
Na zabezpečenie gramatickosti viet generovaných vzormi vyššie v hierarchii je ale potrebné zachovať závislosť vznikajúcich vzorov na kontexte, teda pevných (hoci potenciálne málopočetných) výskytoch konštituentov v ich bezprostrednom okolí ako v pôvodnom modeli, keďže daná kombinácia afixov a význačných slov môže byť používaná aj v rozličných významoch (a kontextoch) a navyše je nutné ošetriť fakt, že nie každá konštrukcia jazyka je viazaná na takéto kombinácie poskytnutím možnosti ľubovoľného slova na vybraných pozíciách vzoru. Napriek takejto závislosti použitie dostatočne voľného (respektíve nie tematicky ohraničeného) vstupného korpusu nutne vedie k odklonu generalizácií (nevidených viet) od sémantickej vierohodnosti, alebo k vzniku príliš veľkého počtu podobných a prakticky nevyužiteľných vzorov (čo nie je problémom pri akvizícii jazyka ľuďmi). Zároveň sa možno domnievať, že tvorba gramatiky aj sémanticky bezchybných konštrukcií a tried ekvivalencie je spojená s pragmatickým využívaním a následným revidovaním nevyužiteľných vzorov.

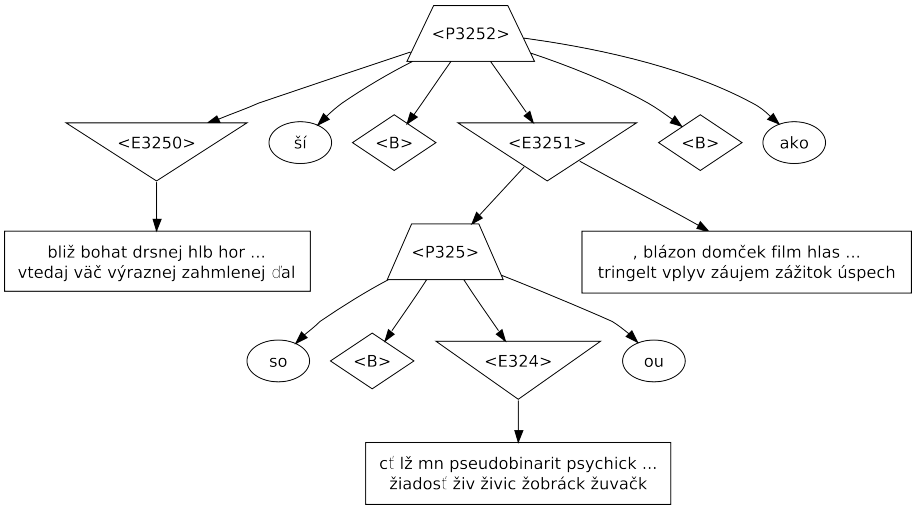
Na vytvorenie obrazu o výsledkoch behu algoritmu na danom korpuse záver tejto podkapitoly bez komentára patrí vybraným príkladom získaných vzorov.

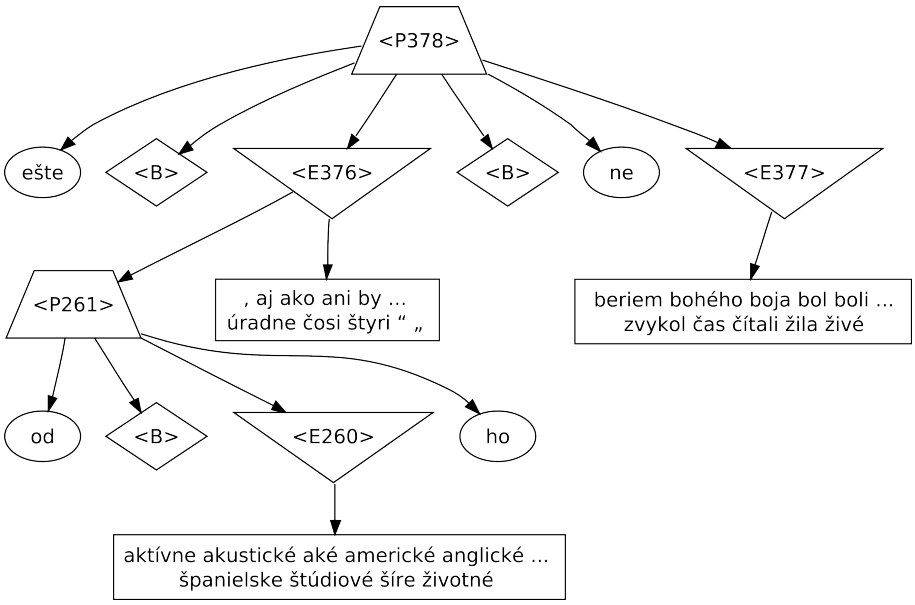


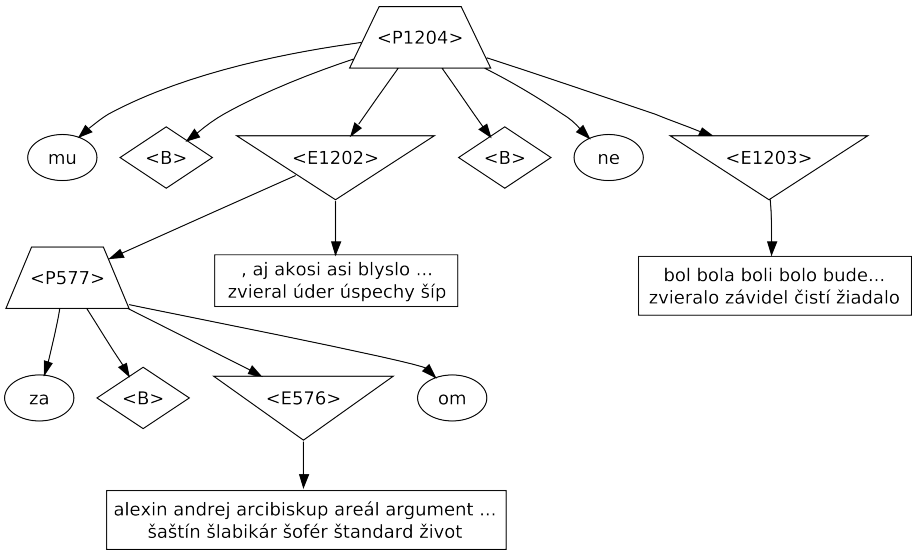


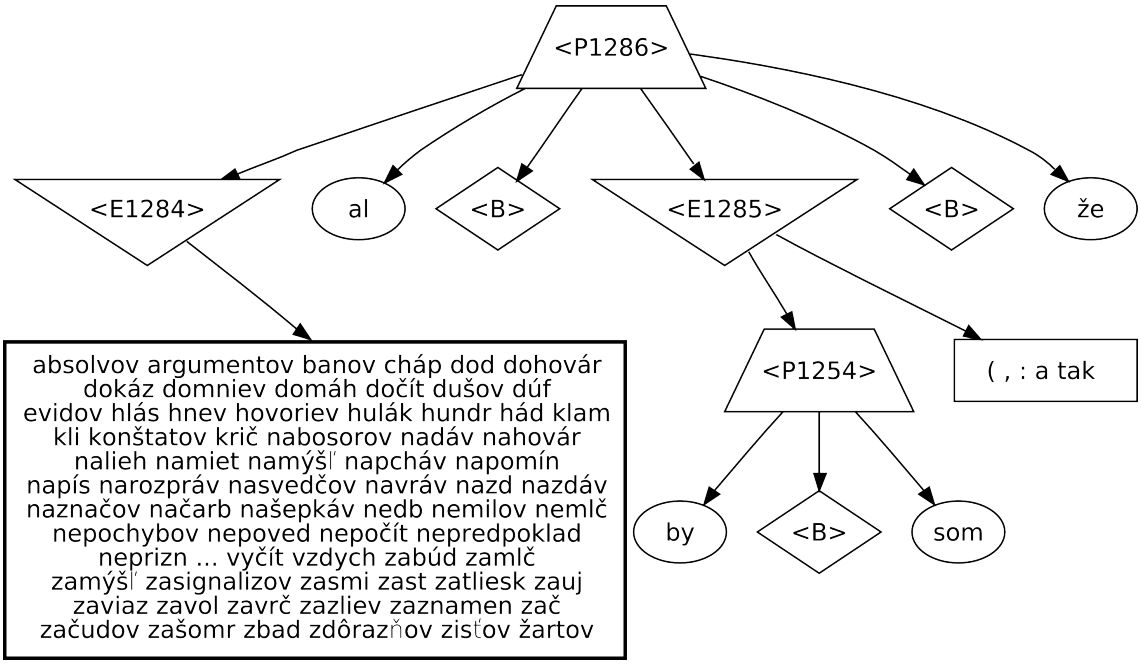


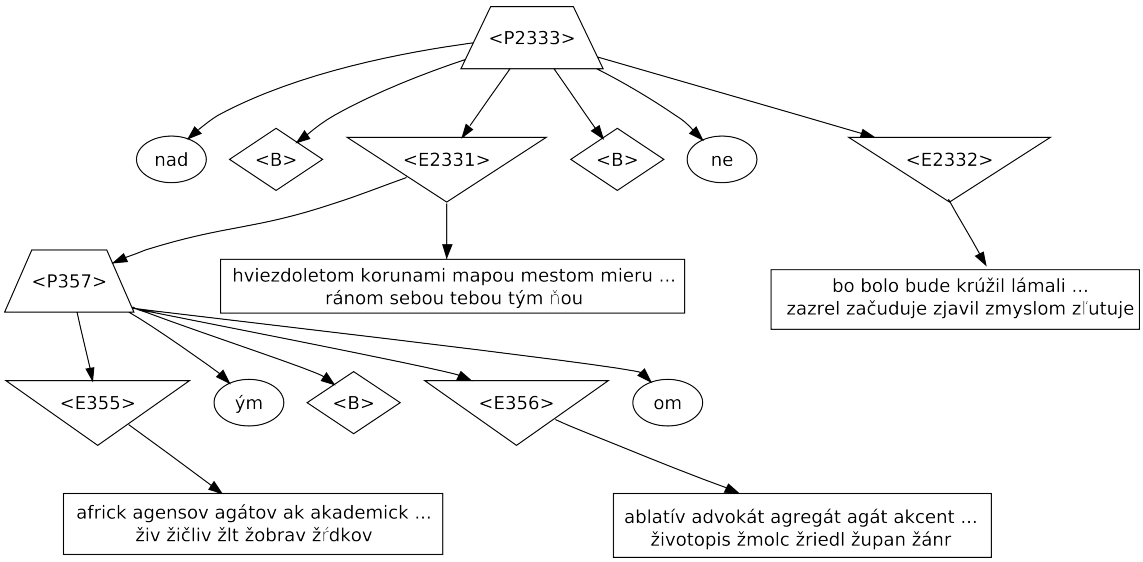


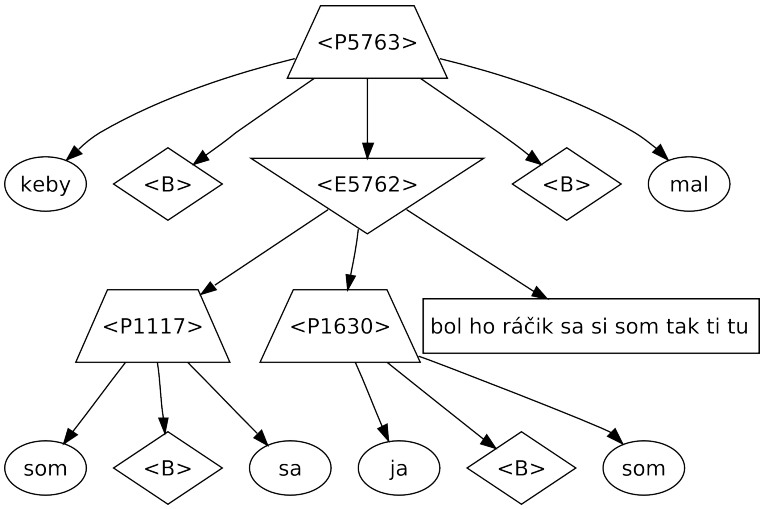

<P3252>
<E3250>
ší
<B>
<E3251>
<B>
ako
bliž bohat drsnej hlb hor ...
vtedaj väč výraznej zahmlenej ďal
<P325>
, blázon domček film hlas ...
tringelt vplyv záujem zážitok úspech
so
<B>
<E324>
ou
cť lž mn pseudobinarit psychick ...
žiadosť živ živic žobrácк žuvačk

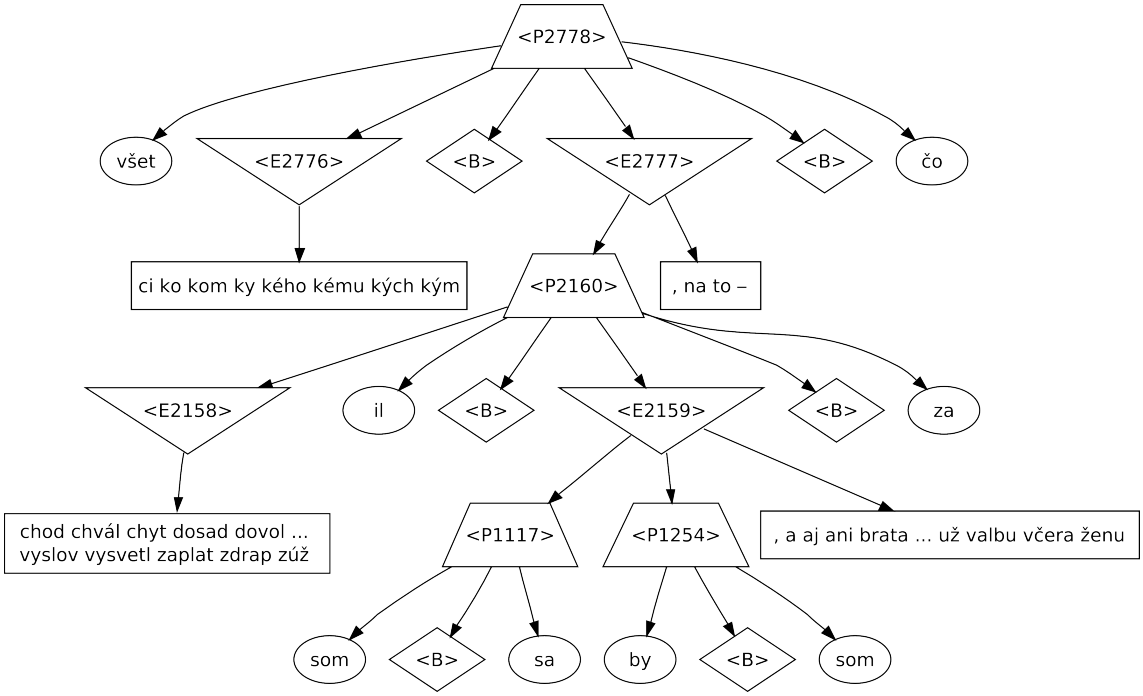

<P378>
ešte
<B>
<E376>
<B>
ne
<E377>
<P261>
, aj ako ani by ...
úradne čosi štyri “ „
beriem bohého boja bol boli ...
zvykol čas čítali žila živé
od
<B>
<E260>
ho
aktívne akustické aké americké anglické ...
španielske štúdiové šíre životné

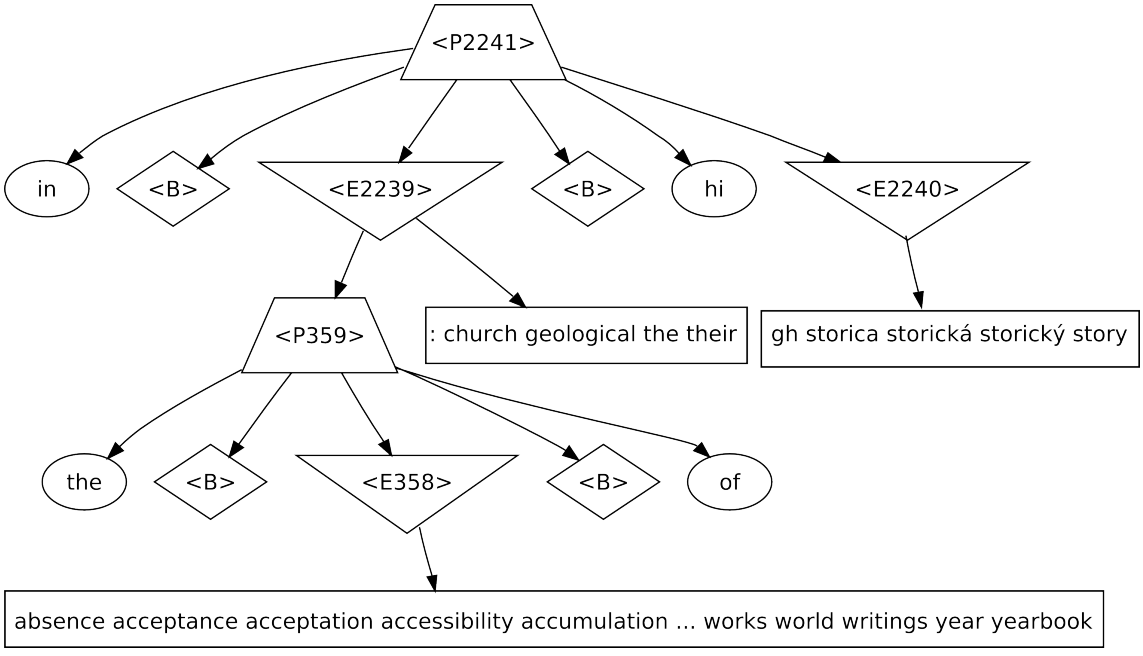

<P1204>
mu
<B>
<E1202>
<B>
ne
<E1203>
<P577>
, aj akosi asi blyslo ...
zvieral úder úspechy šíp
bol bola boli bolo bude...
zvieralo závidel čistí žiadalo
za
<B>
<E576>
om
alexin andrej arcibiskup areál argument ...
šaštín šlabikár šofér štandard život


<P1286>
<E1284>
al
<B>
<E1285>
<B>
že
absolvov argumentov banov cháp dod dohovár dokáz domniev domáh dočít dušov dúf evidov hlás hnev hovoriev hulák hundr hád klam kli konštatov krič nabosorov nadáv nahovár nalieh namiet namýšľ napcháv napomín napís narozpráv nasvedčov navráv nazd nazdáv naznačov načarb našepkáv nedb nemilov nemlč nepochybov nepoved nepočít nepredpoklad neprizn ... vyčít vzdych zabúd zamlč zamýšľ zasignalizov zasmi zast zatliesk zauj zaviaz zavol zavrč zazliev zaznamen zač začudov zašomr zbad zdôrazňov zisťov žartov
<P1254>
(, : a tak
by
<B>
som


<P2333>
nad
<B>
<E2331>
<B>
ne
<E2332>
<P357>
hviezdoletom korunami mapou mestom mieru ... ránom sebou tebou tým ňou
bo bolo bude krúžil lámali ... zazrel začuduje zjavil zmyslom ľutuje
<E355>
ým
<B>
<E356>
om
africk agensov agátov ak akademick ... živ žičliv žlt žobrav žŕdkov
ablatív advokát agregát agát akcent ... životopis žmolc žriedl župan žánr


<P5763>
keby
<B>
<E5762>
<B>
mal
<P1117>
<P1630>
bol ho ráčik sa si som tak ti tu
som
<B>
sa
ja
<B>
som


<P2778>
všet
<E2776>
<B>
<E2777>
<B>
čo
ci ko kom ky kého kému kých kým
<P2160>
, na to –
<E2158>
il
<B>
<E2159>
<B>
za
chod chvál chyt dosad dovol ...
vyslov vysvetl zaplat zdrap zúž
<P1117>
<P1254>
, a aj ani brata ... už valbu včera ženu
som
<B>
sa
by
<B>
som


<P2241>
in
<B>
<E2239>
<B>
hi
<E2240>
<P359>
: church geological the their
gh storica storická storický story
the
<B>
<E358>
<B>
of
absence acceptance acceptation accessibility accumulation ... works world writings year yearbook

6 Záver

Cieľom tejto práce bolo modifikovať a implementovať model Adios tak, aby bol použiteľný na korpuse (cielene Slovenského jazyka) rozdelenom na jednotlivé znaky a bez predchádzajúcej znalosti význačných morfém daného jazyka dokázal vyhľadať v korpuse obsiahnutú informáciu o väzbách slov vyjadrených ich afixami.

Na prezentovaných výsledkoch sa ukázalo, že metóda hľadania postupností ciest so zhodami v postupnostiach znakov medzi hranicami niektorých tokenov alebo v ich okolí dokáže význačné morfémy identifikovať a využiť vo vzoroch konštruovanej gramatiky.

Na zabezpečenie dodržiavania vzdialených väzieb medzi frázami ale tento druh informácie nepostačuje, preto je pri vyhľadávaní korektných konštrukcií na úrovni slovosledu nutné rozdeľovať jednotlivé inštancie prezentovaných nízkoúrovňových vzorov podľa ich vonkajšieho kontextu ako v pôvodnom modeli, pričom rozoznanie vhodných tried kontextov je vzhľadom na variabilitu obsahu a chronický problém nízkeho zastúpenia celkom rovnakých fráz vo všeobecnejšom korpuse problémom hodným ďalšieho skúmania.

7 Bibliografia

Clark, A. S. 2001. Unsupervised language acquisition: Theory and practice. PhD thesis, University of Sussex. [s.n.] 2001.

Edelman, S., Solan, Z., Ruppin, E., Horn, D. 2004a. Learning syntactic constructions from raw corpora. In Proc. of the 29th Boston University Conference on Language Development, Boston, Mass., Nov. 2004.

Edelman, S., Solan, Z., Horn, D, Ruppin, E. Rich syntax from a raw corpus: Unsupervised does it. In Syntax, Semantics and Statistics Workshop of NIPS–2003. 2003.

Edelman, S., Solan, Z., Horn, D, Ruppin, E. 2004b. Bridging computational, formal and psycholinguistic approaches to language. In Proceedings of the Twenty–sixth Annual Conference of the Cognitive Science Society. Chicago. 2004.

Eguchi, S., Copas, J. 2006. Interpreting Kullback–Leibler divergence with the Neyman–Pearson lemma. In J. of Multivariate Analysis 9, vol. 97. San Diego: Academic Press 2006.

Firth, J. R. 1957. A synopsis of linguistic theory 1930–1955. In Studies in Linguistic analysis. Oxford: Philological Society 1957, p. 1–32.

Gold, E. M. 1967. Language identification in the limit. In Information and Control 10. 1967, p. 447–474.

Goldberg, A. E. 2003. Constructions: A new theoretical approach to language. In Trends in Cognitive Sciences 7. London: Trends 2003, issue 5, p. 219–224.

Krenn, B., Samuellson, C. 1997. The linguist's guide to statistics: Don't panic. [online] http://nlp.stanford.edu/fsnlp/dontpanic.pdf. [s.n] 1997.

Kunik, V., Solan, Z. Edelman, S. 2005. Motif extraction and protein classification. In Fourth International IEEE Computer Society Computational Systems Bioinformatics Conference (CSB 2005), Stanford. IEEE Computer Society 2005, p. 80–85.

Manning, C. D., Schütze, H. 1999. Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press 1999.

Roberts A., Atwell E. 2002. Unsupervised grammar inference systems for natural language. Technical Report 2002.20. Leeds: School of Computing, University of Leeds 2002.

Wolff, J. G. 1988. Learning syntax and meanings through optimization and distributional analysis. In Levy, Y., Schlesinger, I. M., Braine, M. D. S. Categories and Processes in Language Acquisition. Hillsdale, New Jersey: Lawrence Erlbaum Associates 1988.

Sag, I. A., Wasow, T. 1999. Syntax theory: A formal introduction. Stanford: Center for the Study of Language and Information 1999.

Solan, Z., Ruppin, E., Horn, D., Edelman, S. 2003a. Automatic acquisition and efficient representation of syntactic structures. In Becker, S., Thrun, S., Obermayer, K. Advances in Neural Information Processing Systems 15. Cambridge, Mass.: MIT Press 2003, p. 90–98. Prístupné online na http://adios.tau.ac.il.

Solan, Z., Horn, D., Ruppin, E., Edelman, S. 2004. Unsupervised context sensitive language acquisition from a large corpus. In Thrun, S., Saul, L. Schölkopf, B. Advances in Neural Information Processing Systems 16. Cambridge, Mass.: MIT Press 2004.

Solan, Z., Horn, D., Ruppin, E., Edelman, S. 2003b. Unsupervised efficient learning and representation of language structure. In Alterman, R., Kirsh, D. Proc. 25th Conference of the Cognitive Science Society, Hillsdale, New Jersey: Lawrence Erlbaum Associates 2003.

A Programová príloha práce

Na priloženom CD nosiči sa nachádza použitá implementácia algoritmu Amigos vo forme projektu pre vývojové prostredie NetBeans 4.0 na platforme J2SE od verzie 5.0, voľne stiahnuteľnej na internetovej doméne http://java.sun.com. Na rovnakej adrese je možné stiahnuť aj potrebný interpreter jazyka java pre rôzne počítačové systémy.

Odporúčaný spôsob spustenia programu je pomocou jednoduchého skriptu run v adresári bin. Program loguje vybrané informácie do nového súboru vytvoreného každým spustením programu v podadresári log súčasného adresára; v prípade neexistencie tohto adresára alebo podobnej chyby sa na toto upozorní prostredníctvom nefatálnej výnimky v programe. Program sa taktiež pokúsi načítať konfiguračný súbor amigos.properties v pracovnom adresári. Do tohto súboru možno zapísať parametre algoritmu, ktorých identifikátory a prednastavené hodnoty sú uvedené v priloženom zdrojovom kóde programu (najmä triedy amigos.paq.PaQAlgorithm a amigos.paq.LWord).

Ako prvý parameter sa programu zadá príkaz. Dôležitými príkazmi sú build na vytvorenie iniciálneho grafu z textových súborov, run na samotný beh algoritmu a extract na výpis získanej gramatiky a iných informácií z grafu. Pomocou príkazu help možno získať informácie o všetkých príkazoch a ich krátky popis.

Na následnú prácu s gramatikou sa v adresári nachádzajú aj moduly pre (aj interaktívny) interpreter jazyka Python (stiahnuteľný pod OSI–certifikovanou licenciou na http://www.python.org). Hlavným modulom je grammar.py, kde sa pomocou metód triedy Grammar dajú získavať informácie o gramatike, generovať vety a merať zhoda s nimi. Na úpravu súborov získaných príkazom extract Amigosu do podoby spracovateľnej týmto modulom postačí kombinácia príkazov štandardných operačných systémov grep a sed, ktorú možno nájsť v skrpite ext2grr v adresári bin. Ďalšia trieda modulu, GrammarGenerator sa v spojení s modulmi montecarlo a lexicon využila pri konštruovaní náhodných gramatík.