SLOVENSKÁ TECHNICKÁ
UNIVERZITA V BRATISLAVE
FAKULTA ELEKTROTECHNIKY
A INFORMATIKY

Daniel Hromada

Autoreferát dizertačnej práce

EVOLUČNÉ MODELY ONTOGENÉZY LINGVISTICKÝCH KATEGÓRIÍ: ŠTYRI SIMULÁCIE

na získanie akademickej hodnosti doktor (philosophiae doctor, PhD.)

v doktorandskom študijnom programe: **Kybernetika**

v študijnom odbore 9.2.7. Kybernetika

Miesto a dátum: Berlín, 9.8.2016

SLOVENSKÁ TECHNICKÁ UNIVERZITA
V BRATISLAVE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Daniel Hromada

Autoreferát dizertačnej práce

EVOLUČNÉ MODELY ONTOGENÉZY LINGVISTICKÝCH KATEGÓRIÍ : ŠTYRI SIMULÁCIE

na získanie akademickej hodnosti doktor (philosophiae doctor, PhD.)

v doktorandskom študijnom programe:

9.2.7. Kybernetika

Miesto a dátum: Berlín, 9.8.2016

Dizertačná práca bola vypracovaná v externej forme doktorandského štúdia

Na Ústav Robotiky a Kybernetiky, Fakulta Elektrotechniky a Informatiky, Slovenská Technická Univerzita v Bratislave

École Doctorale Cognition Langage Interaction, Université Paris 8

Predkladateľ: Daniel Hromada

Universität der Kunste, Grunewaldstrasse 2, 10823, Berlín, Nemecko

Školiteľ: doc. Ing. Ivan Sekaj, PhD.

prof. Charles Tijus

Oponenti: doc. RNDr. Mária Markošová, PhD.

Katedra aplikovanej informatiky FMFI UK, Mlynská dolina 824 48 Bratislavq

Youssef Chahir, Associate Professor, HDR

GREYC Laboratory CNRS UMR 6072, Department of Computer Science

University of Caen Lower-Normandy

Autoreferát bol rozoslaný: 9.8.2016

Obhajoba dizertačnej práce sa koná: 5.9. 2016 o 10:00 h.

Na Ústav Riadenia a Kybernetiky, FEI STU

Ilkovičova 3, 812 19 Bratislava

Obsah

1 Úvod

Dizertačná práca „Evolučné modely ontogenézy lingvistických kategórií“ (ďalej len „Dizertácia“) je výsledkom kombinácie štúdijného programu Kybernetika akreditovanom na Fakulte Elektrotechniky a Informatiky Slovenskej Technickej Univerzity a štúdia kognitívnej psychológie realizovaného na doktorandskej škole Ecole Doctorale Cognition Langage Interaction ktorá je pridružená k Univerzite Paríž 8 St. Denis, ktorá je od roku 2014 zakladajúcim členom asociácie vysokých škôl združených pod názvom Paris Lumières.

Jedná sa teda o dizertáciu interdisciplinárnu, o dizertáciu ktorej prvotným cieľom a zmyslom je podnietiť racionálnu diskusiu medzi kognitívnymi vedami orientovanými ako humanitne (psychológia, jazykoveda, filosofia, didaktika), tak matematicko - informačne (informatika, kybernetika, strojové učenie).

Dizertácia nadväzuje na autorovu takmer 300-stranovú monografiu „Konceptuálne Základy – Intramentálna Evolúcia a Ontogenéza Detskej Reči“ (Hromada, 2015a) v ktorej sa autor pokúsil bližšie objasniť a popísať teoretické a empirické východiská (viz. nižšie) z ktorých vychádzal pri vývoji a ladení štyroch programov. Tých štyroch programov ktoré možno považovať za „výpočtové“ jadro ako Dizertácie samotnej, tak celého cyklu Propedeutica Didactica ktorého je Dizertácia druhým zväzkom.

Každý zo spomenutých štyroch programov je popísaný v osobitnej kapitole. Každý popis má podobu vedeckého článku ktorý sa dá čítať samostatne a bol, alebo bude, odpublikovaný ako samostatná vedecká publikácia. Každá z prvých štyroch kapitol tiež obsahuje tzv. „všeobecný úvod“ a „všeobecný záver“ ktoré usiluje o zasociovanie jednotlivých kapitol medzi sebou. Napokon je dizertácia uzatvorená kapitolou „Summa“ ktorej ašpiráciou je zosúladenie jednotlivých čiastkových motívov s takzvanou Teóriou Intramentálnej Evolúcie.

V súlade s dohodou uzatvorenými v roku 2012 medzi Slovenskou Technickou Univerzitou a Univerzitou Paríž 8 je Dizertácia písaná v anglickom jazyku.

2 Východiská

2.1 Teoretické východisko

Ústredné teoretické východisko Dizertácie je vymedzené hypotézami „učenie je formou evolúcie“, „učenie možno simulovať pomocou evolučných výpočtov“, „učenie ľudskej reči možno simulovať pomocou evolučných výpočtov“ resp. „spôsob akým sa ľudské deti učia reči svojich matiek možno simulovať pomocou evolučných výpočtov“ (Hromada, 2015a).

Analógia medzi učením a evolúciou nieje žiadne *nóvum* a je v mierne pozmenených podobách prítomná v prácach zakladateľov disciplín akými sú napr. evolučná teória poznania (Campbell, 1974), genetická psychológia (Piaget, 1974) či - do istej miery - aj memetická teória (Dawkins, 1976; Blackmore, 2000). A jedná sa vskutku o analógiu plodnú, ktorá na jednej strane umožnila nositeľovi Nobelovej ceny Konradovi Lorenzovi premostiť vedu zoologickú s vedou kybernetickou (Lorenz, 1973) , na strane druhej však tiež umožnila inému nositeľovi Nobelovej ceny G.Edelmanovi postulovať tzv. „teóriu neurálneho darwinizmu“ (Edelman, 1987).

Všetky vyššie uvedené disciplíny a paradigmy možno vnímať ako konkrétne inštancie doktríny tzv. „univerzálneho darwinizmu“ ktorá tvrdí, že biologicko-geneticko evolučný proces - tj. proces ktorého materiálnym substrátom sú vlákna ribonukleových kyselín (Watson & Crick, 1953) - je len jedným z mnohých evolučných procesov ktoré prebiehajú nielen v prírode či v spoločnosti, ale možno aj v *(intra*) jednotlivých ľudských mysliach (*mentes*). A práve popis, či aspoň náznak, výpočtového dôkazu existencie takého tretieho, *intramentálneho* evolučného procesu, je ultimátnym cieľom o ktorý sa Dizertácia usiluje.

2.2 Empirické východisko

Jeden spôsob ako možno preukázať platnosť určitej hypotézy spočíva v metodickom pozorovaní a zaznamenávaní určitých empirických skutočností. Že sa jedná o spôsob pre vedu užitočný je zrejmé už z toho, že takmer všetky významné osobnosti západnej vedy - Aristotelom či Galileom počnúc a Darwinom (Darwin, 1859) či Mendelom končiac - iniciovali výstavbu svojich konceptuálnych systémov práve fázou čistého pozorovania.

Ináč tomu nieje ani v prípade vedy ktorá sa usiluje o pochopenie vývoja jazykových a duševných funkcií ľudských detí. Mnohí vedci (napr. Brown, 1958; Labov&Labov, 1978; Tomasello, 2009) stojaci u základov tejto vednej disciplíny, nazývanej aj vývojová psycholingvistika, zahájili svoju prácu práve kvalitatívnymi pozorovaniami rečových prejavov svojich vlastných detí.

A inak tomu nebolo ani v prípade autora Dizertácie, ktorý na stranách 168-206 Konceptuálnych Základov (Hromada, 2015a) niekoľko tuctov konkrétnych prípadov, kedy verbálny prejav autorovej dcéry I.M. preukázal náznaky toho, že v mysli I.M. možno prebieha proces nielen intramentálny, ale aj evolučný.

Niečo podobné naznačili následne (Ibid., pp.207-233) aj kvantitatívne, data-mining pripomínajúce analýzy tisícov textových transkriptov zahrnutých vo verejne dostupnom korpuse Child Language Data Exchange System (CHILDES, MacWhinney&Snow, 1991). Povedané konkrétnejšie: analýzy korpusu CHILDES naznačili že verbálne interakcie medzi malými deťmi a ich rečovým prostredím (ktoré je určené najmä špecifickou rečou matiek, tzv. „motherese“ alebo „materština“) sú plné javov a štruktúr ktoré možno charakterizovať ako výsledky pôsobenia rozličných inter- či intra- mentálnych variačných operátorov.

Za špeciálne prípady takýchto operátorov *variácie* možno považovať „hravosť“, „tvorivosť“, „vymýšľavosť“, „zabúdanie“ či iné. Prirodzená tendencia detí a matiek mnohé jazykové štruktúry neustále opakovať je zas interpretovaná ako forma replikácie: *repetícia je replikácia*. Napokon je tiež naznačené že vysoká výpočtová zložitosť mnohých problémov spojených s interiorizáciou jazykových pravidiel a kategórií je vyriešená jednoducho tak, že rečový input je principiálne predspracovaný matkou, učiteľom či starším rovesníkom ktorý vo vzťahu k učiacemu sa dieťaťu plnia rolu tzv. „výpočtovej Oracle“ (Turing, 1939) resp. „minimálne adekvátneho učiteľa“ (Clark, 2010). *Selekcia* tak v istom zmysle prebieha nielen v agentovi ktorý sa učí, ale v prvom a neposlednom rade aj v učiteľovi ktorý agenta vystavuje dátam, *z ktorých* sa učí.

3 Ciele Dizertácie

Ultimátnym cieľom Dizertácie bolo predložiť dôkaz „*ex computatio atque simulatio*“ pre hypotézu ktorá tvrdí že „*proces učenia sa materskému jazyku možno simulovať pomocou evolučného algoritmu*“.

Už v prvých fázach prípravy Dizertácie sa však ukázalo že uvedený „ultimátny“ cieľ nieje dosiahnuteľný v prípade že najsamprv nebudú dosiahnuté isté ciele teoretické. Ako obzvlášť závažný sa ukázal byť problém súvisiaci s odpoveďou na otázku:

„Čo sú to „rečové kategórie“ a ako ich možno formalizovať tak, aby sa dali lokalizovať pomocou evolučných výpočtov?“

Zodpovedanie tejto otázky sa ukázalo ako obzvlášť závažné najmä z toho dôvodu, že bez robustnej definície pojmu „kategória“[1] nemožno očakávať robustnú definíciu pojmov ako „gramatické pravidlo“ či „gramatika“. Nezávisle od toho či totiž zvolíme prístup založený na generatívnych gramatikách (Chomsky, 1957), gramatických systémoch (Jimenez-Lopez, 2000), jazykových kolóniách (Kelemen, 2004) či na iných, viac konštrukčne-orientovaných gramatikách (Fillmore et al., 1988; Lakoff, 1990; Tomasello, 2009) , stále budeme konfrontovaný s nutnosťou charakterizovať jednotlivé komponenty rečového prejavu pomocou neterminálnych symbolov a|alebo syntagmaticko-paradigmatických (Saussure, 1916) kategórií.

Alebo, ako naznačujeme inde: „bez znalosti kategórií niet znalosti gramatických pravidiel a bez znalosti gramatických pravidiel možno len ťažko obdržať náležitú a presnú znalosť kategórií“ (Hromada, 2014).

A práve vytvorenie evolučného algoritmu schopného vyvodiť z textu T gramatiku G jazyka J v ktorom je T napísaný bolo prvým problémom ktorý nás následne priviedol aj k ostatným cieľom a témam Dizertácie. Tento problém, nazývaný aj *inferencia či indukcia gramatiky* bol doposiaľ len veľmi zriedkavo riešený pomocou evolučných výpočtov (Aycinena et al. 2003, Smith & Witten, 1995) Naopak, prevládajú modely čisto symbolické (Wolff, 1988), konekcionistické (Elman, 1993) či grafové (Solan et al., 2005).

Z tohoto dôvôdu možno za „najdôležitejší“ cieľ celej Dizertácie považovať:

„Vývoj a zverejnenie evolučného algoritmu schopného vyindukovať gramatické pravidlá z čistého textu“

pričom pod pojmom „text“ nerozumieme množiny umelo vygenerovaných sekvencií symbolov. Naopak, v prípade všetkých inputov ktoré v štyroch simuláciách Dizertácie spracuvávame predpokladáme *a priori* že sú nosičom istej informácie a kódujú určitý zmysel .

1 Termín „kategória“ je v rámci Dizertácie používaný prakticky ako synonymum pre termín „trieda“ (class) . Termín „pojem“, resp. „koncept“ je zas používaný v zmysle ktorý je prakticky ekvivalentný s pojmom ktorý označujeme výrazom „sémantická kategória“, resp. „sémantická trieda“.

Či už použitím slova „učenie“ alebo „pravidlo“, všetky takto definované ciele implicitne odkazujú na *zovšeobecňovaciu schopnosť*. Pointou generatívneho pravidla je jeho schopnosť vygenerovať aj to čo bolo doposiaľ nevidené; zmyslom učenia je vytvoriť si model ktorý tej-čo-sa-učí umožní správne zatriediť aj to, čo bolo doposiaľ neznalosťou zastrené.

V tomto kontexte je zrejmé, že evolučný algoritmus ktorého zostrojenie - a zverejnenie - možno vnímať ako najdôležitejší cieľ našej Dizertácie by mal byť považovaný za istý špecifický variant strojového učenia. Keďže vyššie uvedený problém „indukcie gramatiky z čistého textu“ - možno chápať ako špecifický prípad strojového učenia „bez učiteľa“ a keďže je všeobecne známe že problémy strojového učenia bez učiteľa sú zväčša ťažšie riešiteľné ako problémy strojového učenia „s učiteľom“, pokúsili sme sa aj my dosiahnúť najprv cieľ čiastkový, popísateľný slovami:

„Ukázať že problém viactrednej klasifikácie – tj. strojové učenie s učiteľom – je riešiteľný pomocou evolučných algoritmov“

predtým ako sme sa odhodlali k dosiahnutiu ciela:

„Ukázať že problém inferencie gramatických pravidiel z čistého textu – t.j. strojové učenie bez učiteľa – je riešiteľný pomocou evolučných algoritmov“

ktorý je, mutatis mutandi, ekvivalentný s „najdôležitejším“ cieľom uvedeným vyššie.

4 Štyri programy

Štyri programy tvoriace jadro Dizertácie zdieľajú následovné charakteristiky:

- ich vstupmi sú množiny znakových reťazcov
- snažia sa nájsť riešenia pre problémy lingvistickej podstaty
- problémy sú riešené pomocou evolučných výpočtov

Všetky štyri programy sú napísané v programovacom jazyku Practical Extraction and Reporting Language (PERL) ktorý je jedným z najstarších voľne dostupných, vysoko-úrovňových skriptovacích jazykov. Ako programovací jazyk ktorý bol vyvinutý jazykovedcom Larrym Wallom na uľahčenie života pre iných jazykovedcov, obsahuje PERL natívnu implementáciu operátorov ktoré umožňujú nesmierne rýchle vykonávanie takých operácií so znakmi ako sú „pattern matching“ alebo znakové substitúcie (Wall, 1994). V tomto ohľade je PERL kanonickejší ako iné vysokoúrovňové jazyky Python či R ktoré len druhotne implementujú tzv. „PERL-kompatibilné regulárne výrazy“ (PCRE) ktoré boli vyvinuté práve pre PERL[2].

Prvý až tretí program tiež implementujú následovné prístupy a ideje:

- vstupné textové dáta sú premietnuté do vektorových 128 alebo 64 rozmerných euklidovských priestorov pomocou prístupov „náhodného indexovania“ (Sahlgren, 2005) resp. „reflektovaného náhodného indexovania“ (Cohen et al., 2010)
- tieto euklidovské priestory sú následne pretransformované do binárnej podoby pomocou algoritmu „ľahko stochastickej binarizácie“ (Hromada, 2014c)

2 PCRE sú, popri regulárnych výrazoch daných normou POSIX, jedným z dvoch hlavných a vývojarskou komunitou akceptovaným štandartov zápisu regulárnych výrazov. Z tohoto dôvodu uvádzame v niektorých výpisoch kódu, resp. v Appendixoch 1 a 2 nie pseudokód ale priamo PERLový kód. Niektoré PERLové konštrukcie prípadne PCRE sú totiž tak komplexné že snaha vyjadriť ich pomocou iného formalizmu by mohla nášmu komunikačnému zámeru viac uškodiť ako pomôcť. C.f. Hromada (2011b) a Hromada (2016b) v ktorých bližšie obhajujeme náš postoj vskratke vyjadriteľný slovami „PCRE sú formalizmy samé o sebe“.

- kategórie sú formalizované ako konvexné regióny binárnych priestorov (tj. Hammingove sféry)
- hľadanie ideálnych konštelácií takto definovaných kategórií je implementované ako evolučná optimalizácia

Programy jedna až tri v ktorých sú tieto ideje zaimplementované nazývame aj „kategórie-indukujúce“ programy. Za účelom práce s euklidovskými vektormi či maticami je využitá knižnica Perl Data Language (PDL) z ktorej dátové štruktúry a operátory sú značne podobné dátovým štruktúram a operátorom známym z jazykov MATLAB či R. Zároveň však treba dodať že *procedúra vykonaná v najvnútornejšom cykle každého vyhodnotenia účelovej funkcie* je implementovaná v jazyku assembler[3]. Toto umožňuje volať priamo inštrukciu POPCNT ktorá je hardwarovo implementovaná na všetkých procesoroch s inštrukčnou sadou SSE4.2. Výsledkom je značné zrýchlenie evolučného vyhľadávania.

Konkrétne rozdiely medzi „kategórie-indukujúcimi“ programami 1-3 sú zobrazené na nižšie uvedenej ilustrácii:

3 http://wizzion.com/thesis/simulation2/popcount.asm

4.1 Nultý program

Nultý program je predstavený v kapitole „Prielom do neznámeho kódu“. Začína sa slovami: „*Kryptológ konfrontovaný s neznámou šifrou je v pozícii podobnej dieťaťu, ktoré sa práve narodilo. Tak ako dieťa je aj kryptológ vystavený novým konšteláciám symbolov a ich čŕt, ktorých význam mu zatiaľ nieje známy*.“

Analógia medzi dieťaťom a kryptológom následne vedie k predstaveniu jednej z mála nerozlúštených kryptologických hádanok súčasnosti: k takzvanému Voynichovmu Rukopisu (dalej len „V.R.“) , 240 stranovému textu napísanému v doposial nerozlúštenom písme a možno aj neznámom – či dokonca už v zaniknuvšom - jazyku. Po krátkom exkurze popisujúcom minulé snahy o rozlúštenie V.R. či niektoré aktuálne poznatky o tomto reálne existujúcom historickom artefakte, je pozornosť čitateľa upriamená na tzv. „*primary mapping*“, tj. na istú pravidelnosť medzi textom a obrazmi vo V.R. obsiahnutými.

Následne je predstavená evolučná výpočtová metóda ktorej cieľom je nájsť čo najoptimálnejší spôsob prepisu zo sady znakov v ktorých je V.R. napísaný, do sady znakov ktoré reprezentujú alfabetu určitého jazyka. Za optimálnejší je pritom považovaný taký spôsob prepisu, ktorý prepíše čo najväčšie množstvo slov uvedených vo V.R., prípadne jeho časti, do slov ktoré sa vyskytujú v istom dopredu danom slovníku. Za globálne optimum - ekvivalentné rozlúšteniu textu - sa v takom prípade považuje situácia kedy sú všetky slová, uvedené vo V.R., konzistentne prepísané do slov uvedených v slovníku.

Vychádzajúc z predpokladov:

- V.R. je zakódovaný monoalfabetickou substitučnou šifrou
- sekcia V.R. nazývaná aj „kalendár“ obsahuje zoznamy krstných mien

sme následne vyvinuli relatívne jednoduchý evolučný algoritmus ktorého vstupom je „kalendár“ na strane jednej a „slovník“ na strane druhej. Bolo vyskúšaných viacero slovníkov a vo všetkých prípadoch preukázal evolučný algoritmu schopnosť adaptovať „kalendár“ na „slovník“. Najlepšie výsledky však boli dosiahnuté v prípadoch kedy „slovník“ obsahoval zoznamy ženských krstných mien s obráteným poradím písmen (tj. písané zľava doprava).

Napr. v jednom z prípadov - kedy bol ako „slovník“ použitý zoznam obsahujúci veľké množstvo foriem slovanských ženských mien, vrátane ich zdrobnelých foriem – boli nájdené konzistentné prepisy pre 240 znakov obsiahnutých v niekoľkých desiatkach „hypoteticky rozkódovaných“ znakových reťazcov. Povzbudivé výsledky boli dosiahnuté aj v prípade keď bol použitý oveľa kratší slovník hebrejských ženských mien.

Považujeme za potrebné dodať že napriek tomu že naša metóda vedie k „počtom hypoteticky rozkódovaných znakových reťazcov“ ktoré v nám známej literatúre nemajú obdobu, si ani zďaleka nedovoľujeme tvrdiť že sa nám podarilo V.R. rozľúštiť. Ak je totiž vôbec niečo ako rozkódovanie V.R. možné istotne sa bude jednať o úsilie ktoré presahuje schopnosti akéhokoľvek ľudského jednotlivca. O úsilie ktoré od vedcov vyžaduje intenzívnu interdisciplinárnu spoluprácu. V prípade že k takej spolupráci dôjde nemožno vylúčiť že nami predstavená „evolučná metóda slovníkového útoku“ by sa mohla ukázať ako užitočná najmä z toho dôvodu, že môže značným spôsobom urýchliť identifikáciu množín substitúcií ktorými môže byť V.R. - či iný zašifrovaný korpus - úspešne prepísaný do zmysluplného textu.

Napriek tomu že problém o ktorého riešenie sa nultý programom usiluje súvisí s problémom „indukcie rečových kategórií“ len veľmi vzdialene bol napokon do Dizertácie zaradený preto, že

- implementuje evolučné vyhľadávanie optimálneho spôsobu prepisu
- ilustruje schopnosť evolučných výpočtov adaptovať jednu sadu reprezentácií (napr. „kalendár“ na inú sadu reprezentácií (napr. „slovník“)

ako aj preto, že sa pre autora jednalo o prakticky prvý v PERLe napísaný evolučný algoritmus.

4.2 Prvý program

Prvý program je prvým z troch programov ktorým možno prisúdiť prívlastok „kategóriu indukujúci“. Jadro kapitoly, v ktorej je prvý program popísaný, je prakticky totožné s článkom „Genetic optimization of semantic prototypes for multiclass document categorization“ (Hromada, 2015b) za ktorý bola autorovi na doktorandskej konferencii Elitech 2015 udelená cena v sekcii „Aplikovaná informatika“.

Algoritmus ktorý je v článku predstavený možno chápať ako špecifický prípad strojového učenia s učiteľom. Algoritmus kombinuje tzv. teóriu prototypov (Rosch & Mervis, 1975; Rosch, 1999) s evolučnou optimalizáciou a prístupmi založenými na redukcii a binarizácii.

Textové dokumenty obsiahnuté v trénovacom korpuse sú najprv premietnuté do 128-rozmerného priestoru pomocou algoritmu „reflektovaného náhodného indexovania“ (Cohen et al., 2010). Tieto euklidovské priestory sú následne pretransformované do binárnej podoby pomocou algoritmu „ľahko stochastickej binarizácie“ (Hromada, 2014c). Črty na základe ktorých sú jednotlivé vektorové reprezentácie vytvárané, sú principiálne určené frekvenciami výskytu jednotlivých slov v dokumentoch obsiahnutých vo vstupnom textovom korpuse. Výstupom tejto prvotnej, pred-optimizačnej fáze algoritmu je množina 128-rozmerných binárnych vektorov: každá črta (napr. výskyt určitého slova v dokumente) rovnako ako každý objekt (napr. textový dokument) je tak na konci tejto fázy charakterizovaný pomocou určitej špecifickej 16-bajtovej sekvencie, tj. „hašu“.

Považujeme za dôležité zdôrazniť že hašovacia funkcia založená na reflektovanom náhodnom indexovaní priradzuje podobným objektom podobné haše (i.e. haše medzi ktorými je nízka Hammingová vzdialenosť). A naopak: objektom ktoré sú rozdielne budú pravdepodobne priradené haše ktorých vzájomná Hammingovská vzdialenosť je relatívne vysoká. Toto je dôsledkom tzv. Johnson – Lindenstraussovej lemmy ktorá tvrdí že „*ak premietneme body vektorového priestoru do náhodne zvoleného podpriestoru dostatočnej dimenzionality, budú vzdialenosti medzi takto premietnutými bodmi približne zachované*“ (Sahlgren, 2005).

Po prvotnej fáze počas ktorej sú jednotlivým objektom priradené ich geometrické reprezentácie nasleduje fáza optimizačná. Tá spočíva vo vyhľadávaní najoptimálnejšej konštelácie prototypov pomocou kanonického genetického algoritmu (Goldberg, 1990). Účelová funkcia vychádza z kognitívne plauzibilnej (Hromada, 2014b) definície tvrdiacej:

Súradnice ideálneho prototypu kategórie K sú v čo najbližšej možnej blizkosti súradníc súcien ktoré sú v K obsiahnuté a v čo najväčšej možnej vzdialenosti od súcien ktoré v K obsiahnuté niesú.

Účelová funkcia prvého programu je principiálne daná rovnicou:

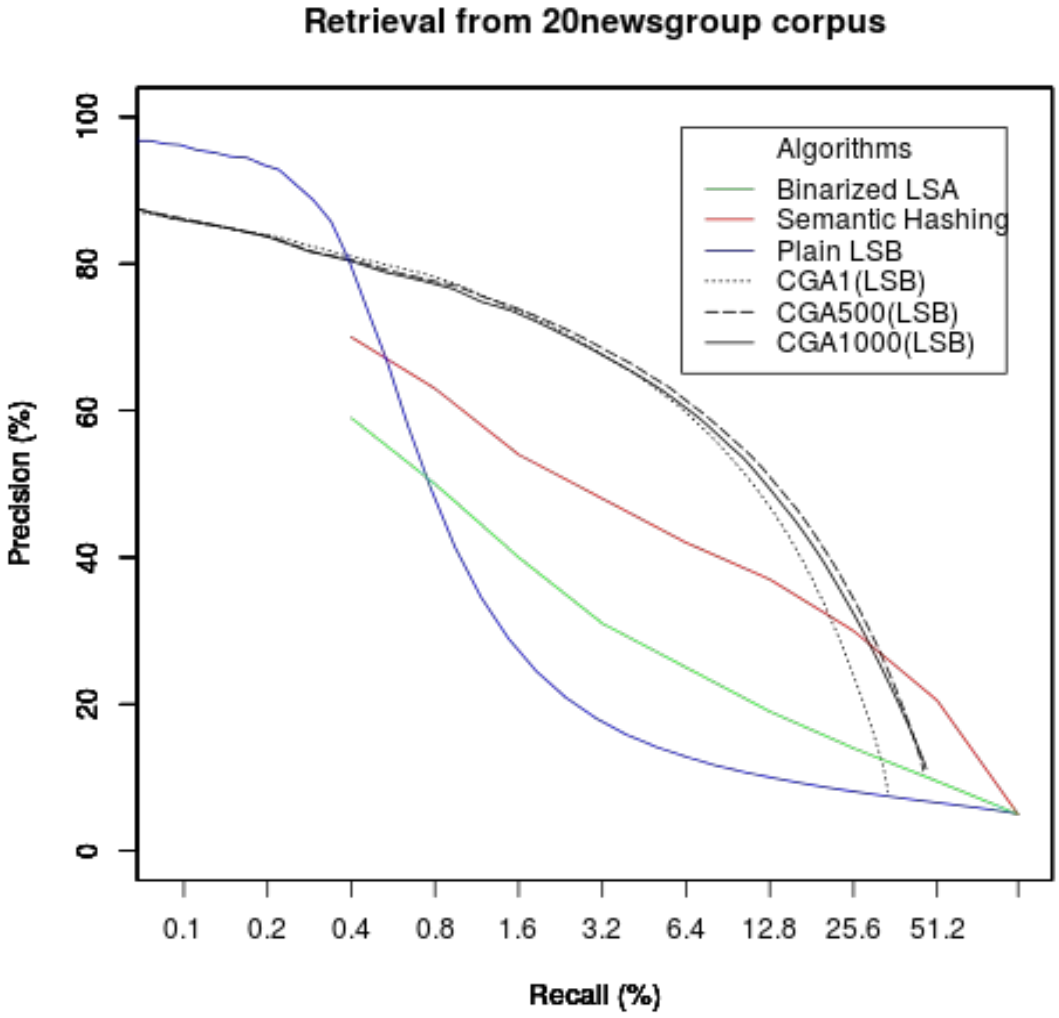
$$F_{CP}(P_K) = \alpha \sum_{t \in C_K} F_{hd}(h_t, P_K) - \omega \sum_{f \notin C_K} F_{hd}(h_f, P_K) \qquad (1)$$

ktorá definuje tzv. klasifikačnú vhodnosť F_{CP} kandidátskeho prototypu P_K kategórie K pomocou rozdieľu medzi súčtom Hammingovských vzdialeností medzi P_K a "vhodnými" (t.j. nachádzajúcimi sa v K) resp. "nevhodnými" (t.j. nachádzajúcimi sa v K) objektami trénovacieho korpusu. Evolučný proces následne možno chápať ako proces ktorý minimalizuje F_{CP} pre prototypy všetkých kategórií:

$$I = \min \sum_{K=0}^{K=|L|} F_{CP}(P_K) \qquad (2)$$

Použiteľnosť algoritmu bola následne vyhodnotená vzhľadom k problému roztriedenia tisícov článkov obsiahnutých v korpuse 20 newgroups do ich náležitých kategórií. Prístup bol porovnaný s technikou „hĺbkového učenia“ známou pod názvom Semantic Hashing (Salakhutdinov & Hinton, 2009), ktorá, podobne ako náš prístup, prioritne operuje s reprezentáciami textových dokumentov majúcich podobu 128-bitových (i.e. 16-bajtových) booleovských vektorov.

Porovnanie výsledkov ukázalo, že v mnohých prípadoch môžu evolúciou identifikované konštelácie sémantických prototypov klasifikovať texty prinajmenšom tak precízne, ako Semantic Hashing.



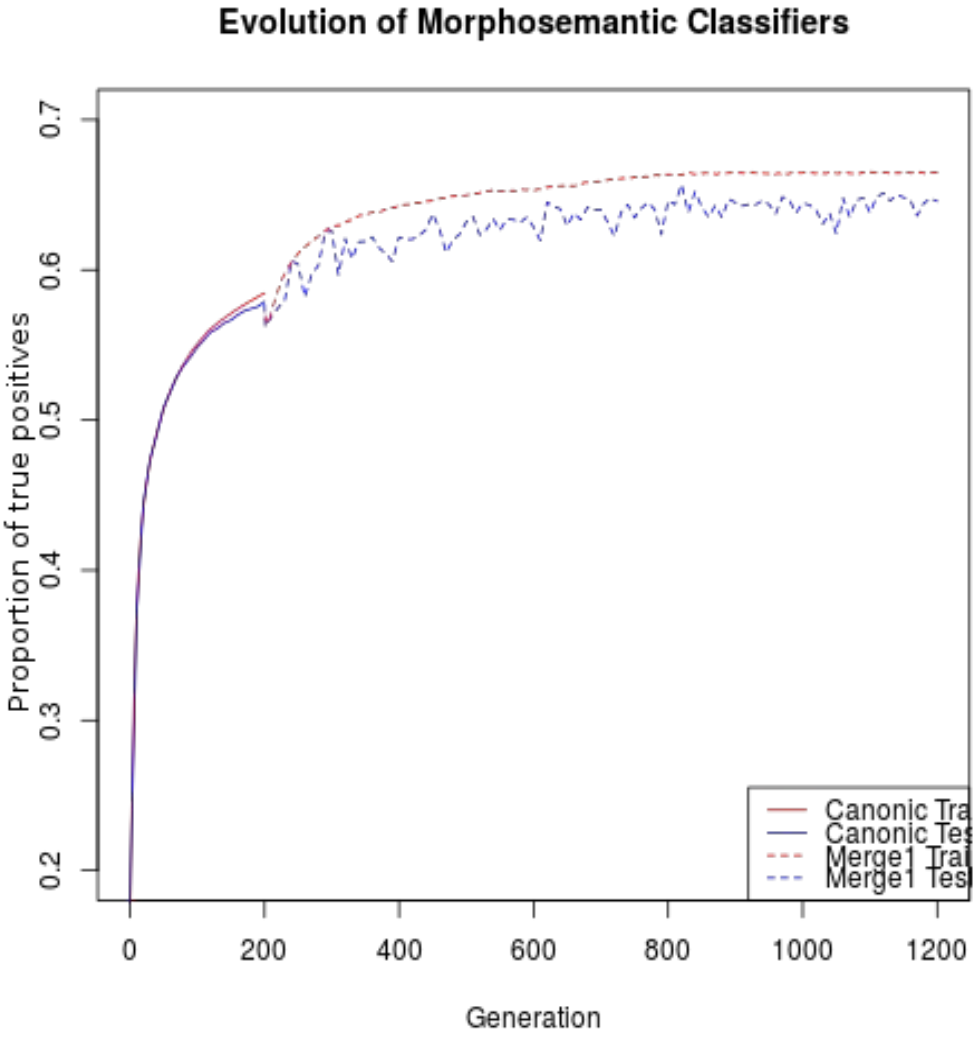
Tieto výsledky naznačujú že evolučného vyhľadávanie konštelácií prototypov môže za určitých okolností viesť k zmysluplným a užitočným riešeniam problému známeho ako strojové učenie viactredneho klasifikátora.

4.3 Druhý program

Druhý program používa evolučné výpočty za účelom vyriešenia problému tzv. "indukcie tvaroslovných kategórií". Celková architektúra je v značnej miere podobná architektúre prvého programu: rovnako ako v prvom programe preto evolučnej optimizácii predchádza "geometrizačná" fáza počas ktorej sú entity vstupného trénovacieho korpusu premietnuté do binárneho priestoru. Podstatné rozdiely medzi prvým a druhým programom tak možno badať najmä v následovných ohľadoch:

- vstupnými dátami sú textové prepisy obsiahnutými v korpuse CHILDES (MacWhinney & Snow, 1991; Brown, 1973) ktorými sa matka prihovárala svojej 2-3 ročnej dcére Eve
- triedy ktorých prototypy sa pokúšame lokalizovať niesú triedami sémantickými (ako tomu bolo v prípade prvého programu), ale triedy morfosémantické alebo tvaroslovné - t.j. triedy ktoré sú v nemalej miere podobné tým čo tradičná jazykoveda nazýva "slovné druhy"[4]
- výskyt každého slova (t.j. "token") je charakterizovaný pomocou troch čŕt: prípona dotyčného slova S, prípona slova ktoré slovu S vrámci daného výskytu predchádza a prípona slova ktoré slovo S vrámci daného výskytu následuje[5]
- účelová funkcia usiluje o dosiahnutie stavu kedy sú geometrické reprezentácie všetkých tokenov obsiahnutých v trénovacom korpuse bližšie k správnym prototypom ako k prototypom nesprávnym
- optimizačný algoritmus $MERGE_1$ možno chápať ako 1-stupňový hierarchický Paralelný Genetický Algoritmus (Sekaj, 2004)

Ukazuje sa že evolučná lokalizácia konštelácií prototypov vedie k zvýšeniu precíznosti päťtriednej morfosémantickej klasifikácie a to ako vo vzťahu k dátam trénovacím, tak testovacím:



4 Dizertácia sa zameriava len na indukciu päť základných morfosémantických tried "substancia", "činnosť", "atribút", "vzťah" a "referencia" ktorými sú aproximované slovné druhy "podstatné mená", "slovesá", "prídavné mená", "predložky" a "zámená".

5 Dôvody prečo sme sa rozhodli pre tak minimalistickú množinu inicializačných čŕt sú bližšie predstavené v prácach Slobin (1973) a Hromada (2014a).

Method	Training corpus	Testing corpus
$CENTROID_{HAMMING}$	455 (42.12%)	412 (40.47%)
$CENTROID_{EUCLIDEAN}$	572 (52.96%)	533 (52.35%)
$MEAN(GA_{CANONIC})$	631 (58.44%)	589 (57.88%)
$MEAN(GA_{MERGE1})$	718 (66.51%)	657 (64.57%)
$FITTEST(GA_{MERGE1})$	772 (71.48%)	699 (68.66%)
MSVM2	781 (72.31%)	736 (72.30%)

Získané výsledky naznačujú že genetická optimalizácia signifikantne zvyšuje trénovaciu aj testovaciu presnosť 5-triedneho klasifikátora. Signifikantné zlepšenie výsledkov je dosiahnuté rozšírením kanonického algoritmu o *nadstavbu* danú algoritmom MERGE: nedá sa preto vylúčiťže využitie inej "paralelnej" architektúry bude viesť k ešte lepším výsledkom. Tak či onak, výsledky dosiahnuté algoritmom MERGE sú v značnej miere porovnateľné s výsledkami ktoré dosiahne 5-triedny klasifikátor založený na tradičnej metóde Support Vector Machine (SVM).

K potešujúcim zisteniam vedie aj bližšia, kvalitatívna, analýza "chýb" ktoré evolučnému morfosémantickému klasifikátoru zabránili v dosiahnutiu 100% presnosti. V nejednom prípade totiž algoritmus odmietol zatriediť do nesprávnej kategórie token, ktorému bol v trénovacích dátach priradený nesprávny label. Napríklad v prípade vety "what are you building here?" priradil algoritmus tokenu "building" správny label "činnosť" (t.j. príčastie slovesa "build") aj napriek tomu že anotátori korpusu nesprávne priradili tokenu label "substantívum". Je miestom pre argument do akej miery možno chápať takéto chovanie ako znak určitej robustnosti nášho algoritmu: je však nutné zdôrazniť že k podobnému odmietnutiu adaptovať sa na chybné vstupné dáta došlo opakovane.

4.4 Tretí program

Tretí program používa evolučné výpočty za účelom riešenia problému "gramatickej indukcie". Cieľom gramatickej indukcie je vyvodiť z vstupného textu T gramatiku (t.j. systém pravidiel) G jazyka J v ktorom je T napísaný. V ideálnom prípade by mala byť vyindukovaná gramatika G schopná vygenerovať aj také vety jazyka J, ktoré neboli v pôvodnom T prítomné. Zároveň by však G nemala byť schopná vygenerovať vety ktoré niesú vetami jazyka J.

Gramatická indukcia je v svojej podstate problémom strojového učenia bez učiteľa: vstupné dáta obsahujú len vety jazyka J. Žiadne dodatočné popisné informácie niesú dodané. Z tohto dôvodu nebolo v prípade tretieho programu možné použiť prístupy založené na lokalizácii prototypov ktoré boli implementované vrámci prvého a druhého programu.

Aj napriek tomu však existuje medzi všetkými tromi "kategórie-indukujúcimi" programami značná podobnosť. Všetky tri totiž principiálne nerobia nič iné ako to, že v istom N-rozmernom binárnom priestore vyhľadávajú konštelácie regiónov spĺňajúce určité kritériá.

Kapitola obsahujúca popis tretieho programu začína formálnymi definíciami základných pojmov. G-Kategória ("geometrizovaná kategória") je definovaná ako sféra lokalizovaná v určitom priestore pomocou súradníc svojho stredu a veľkosťou svojho polomeru. H-Kategória ("Hammingovská kategória") je špeciálnym prípadom G-Kategórie ktorá je vnorená do binárneho (t.j. "Hammingovského") priestoru. N-Schéma je schéma ktorá obsahuje popis N G-Kategórií. Je ukázané že 4-schému obsahujúcu popis štyroch 64-rozmerných H-Kategórií možno zakódovať do binárneho vektora o dĺžke $4 * (64 + \log_2 64) = 4 * 70 = 280$ bitov = 35 bajtov.

Podobné 35 bajtov dlhé binárne vektory sú genotypy jedincov ktorí sú vrámci tretieho programu "vyvíjaní". Dizertácia obsahuje zdrojový kód jednoduchej deterministickej procedúry pomocou ktorej možno genotyp premeniť na fenotyp. Fenotyp má podobu regulárneho výrazu. Aplikáciou fenotypického regulárneho výrazu na vstupný text T možno rýchlo a transparentne zistiť koľko viet z T daný jedinec "matchuje". Počet viet prítomných v T ktoré zodpovedajú danej kandidátskej 4-schéme X nazývame "korpusovou senzitivitou" Y_X a považujeme ju za jednu z dvoch hlavných veličín pre výpočet fitness F(X).

Druhou veličinou je tzv. extenzia E_X udávajúca maximálny možný počet viet ktoré môžu zodpovedať N-schéme X. Extenzia E_X je daná ako geometrický produkt extenzie jednotlivých G-kategórií ktoré sú v X obsiahnuté:

$$E_X = \prod_{i=1}^{N} members(G_i)$$

pričom extenzia jednotlivej kategórie G_i - resp. návratová hodnota funkcie *members* - je daná počtom súcien ktorých vzdialenosť od stredobodu G_i j je menšia ako polomer G_i

Fitness N-schémy X je následne daná ako:

$$F_X = \frac{Y_X * Y_X}{E_X}$$

Je zjavné že maximizácia takejto, alebo podobnej, funkcie bude smerovať optimizačný proces smerom k N-schémam ktoré niesú príliš všeobecné (extenzia je uvedená v deliteľovi) no predsa zodpovedajú čo najväčšiemu možnému počtu viet uvedených vo vstupnom texte (umocnená korpusová senzitivita v delencovi).

Okrem dichotómie genotyp - fenotyp implementuje optimizačné jadro tretej simulácie - algoritmus INDUCTOR - aj ďalšie špecifické prvky:

- mutáciam podliehajú nielen súradnice stredov jednotlivých G-Kategórií, ale aj ich polomery
- nultá populácia nieje vygenerovaná úplne náhodne: polomery všetkých jedincov nultej generácie sú nastavené na hammingovskú vzdialenosť 13
- kríženie prebieha len na určitých pozíciách
- použitím tzv. "stratégie presunu pozornosti" (tzv. re-focusing strategy) je docielené to, že jednotlivé behy evolučnej optimizácie budú konvergovať k doposiaľ neobjaveným N-schémam

Z týchto dôvodov možno algoritmus INDUCTOR považovať za hybrid evolučných stratégií (Rechenberg, 1971) a evolučného programovania (Fogel, 1995).

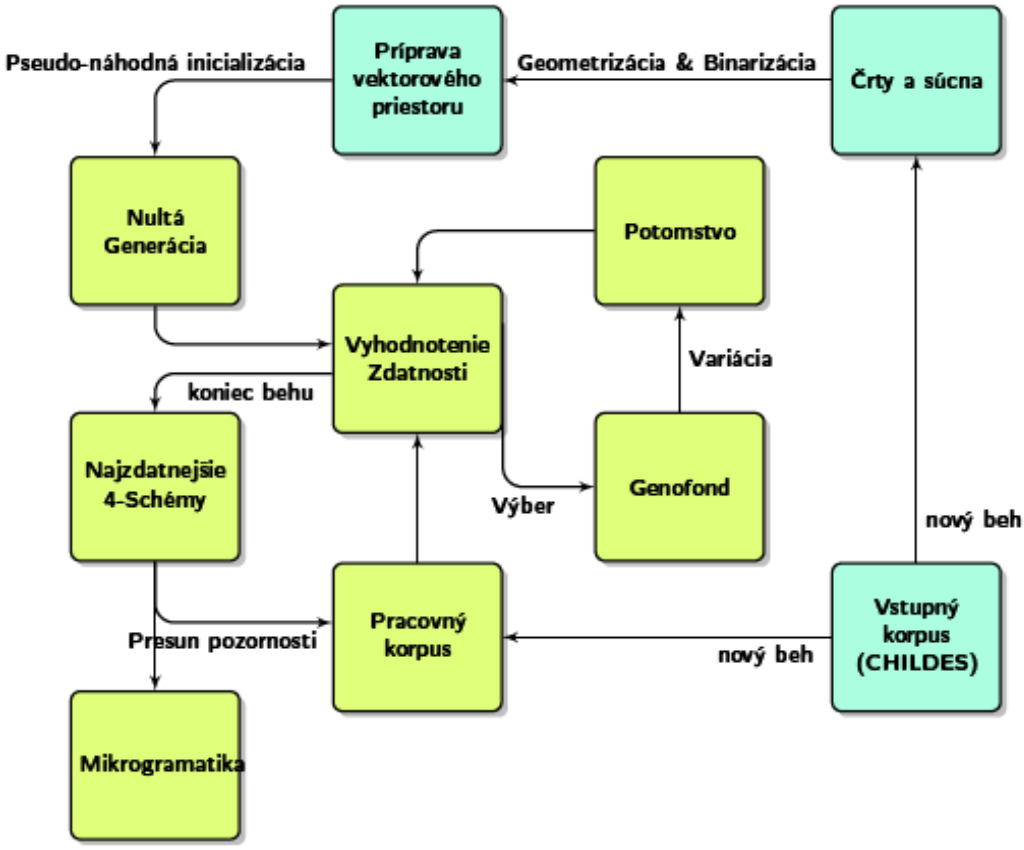
Zmysluplnosť evolučného algoritmu INDUCTOR je ilustrovaná jeho konfrontáciou s 1047 štvorslovnými vetami vyextrahovanými z reči matiek obsiahnutej v anglickej sekcie korpusu CHILDES (MacWhinney & Snow, 1991) . Každé slovo S v týchto vetách prítomné je premietnuté do 64-rozmerného binárneho priestoru na základe čŕt daných slovami ktoré S vo vete predchádzajú resp. po ňom následujú. V takto obdržanom priestore sa následne uskutočňuje 100 „behov“, tj. vzájomne nezávislých evolučných vyhľadávaní 35-bajtov dlhých 4-schém s najvyššou možnou fitness.

Výsledkom je sto 4-schém ktoré dokopy tvoria konštrukčnú mikrogramatiku (Tomasello, 2009) schopnú vygenerovať 176 viet z pôvodných 1047 viet obsiahnutých vo vstupnom texte. Mimo to je však až 82% všetkých identifikovaných 4-schém schopných vygenerovať aj vety ktoré vo vstupnom

texte neboli. Z týchto 82 schém bolo 32 (39%) manuálne identifikovaných ako schémy ktoré produkujú iba syntakticky správne anglické vety. Tak napr. schéma

^(that)(is)(a)(bag|banana|basket|bridge|cherry|cow|gate |horse|kleenex|motorcycle|puzzle|rabbit| raccoon|shoe| spoon|story|timer|tractor)$

dokáže vygenerovať 18 gramaticky korektných anglických viet. Iba päť z týchto osemnástich viet však bolo explicitne uvedených vo vstupnom korpuse.



5 Splnenie cieľov Dizertácie

Prvý cieľ: Zodpovedanie otázky: *„Čo sú to „rečové kategórie“ a ako ich možno formalizovať tak, aby sa dali čo možno najoptimálnejšie identifikovať pomocou evolučných výpočtov?“*

bol splnený následovným spôsobom:

Prvý, druhý aj tretí program implementujú náhľad že kategórie možno formalizovať ako regióny N-rozmerného priestoru (Gärdenfors, 2004). Zmysluplné výsledky sú dosiahnuteľné dokonca[6] aj v prípade že sa jedná o binárne priestory získané metódou náhodnej projekcie.

Kľúčové bolo tiež zistenie že oveľa lepšie výsledky sú dosiahnuté v prípade že kategórie niesú indukované jednotlivo, ale ako súčasť obsiahlejšeho "celku" (paradigmatickej konštelácie prototypov v prípade prvého a druhého programu resp. syntagmatickej 4-schémy v prípade programu tretieho).

6 Prevodom euklidovských súradníc na súradnice vrámci binárneho priestoru nepochybne dochádza k nezanedbateľnej strate informácie. Táto nevýhoda je však v istom zmysle vybalancovaná ľahkosťou a rýchlosťou následnej optimalizačnej fázy *v jadre* ktorej *sa vždy nachádza výpočet vzdialenosti*. Je tomu tak preto,že výpočet hammingovskej je zväčša oveľa menej nákladný ako výpočet skalárneho súčinu medzi dvomi euklidovskými vektormi.

Druhý cieľ: *„Ukázať že problém viactrednej klasifikácie – tj. strojové učenie s učiteľom – je riešiteľný pomocou evolučných algoritmov“*

bol splnený následovným spôsobom:

Prvá simulácia ukázala že textové dokumenty možno do 20 sémantických tried roztriediť pomocou evolučnej lokalizácie konštelácií prototypov. Získané výsledky sú porovnateľné s výsledkami metódy hĺbkového strojového učenia nazývanej Semantic Hashing.

Druhá simulácia naznačila že podobný evolučný proces možno úspešne využiť pri triedení slov do 5 základných slovných druhov. Získané výsledky sú porovnateľné s prístupom metódy strojového učenia známej ako multiclass Support Vector Machine.

Napriek tomu že účelová funkcia sa vyhodnocovala len vzhľadom k trénovacím dátam, prejavili sa pozitíva evolučnej adaptácie aj v zlepšení výsledkov pri triedení dát testovacích. Ináč povedané: evolúciou vyindukované klasifikačné mechanizmy disponujú zovšeobecňovacou schopnosťou. Z tohto dôvodu si myslíme že modely implementované v prvom a druhom programe možno považovať za špecifické prípady strojového učenia s učiteľom.

Tretí cieľ*: „Ukázať že problém inferencie gramatických pravidiel z čistého textu – t.j. strojové učenie bez učiteľa – je riešiteľný pomocou evolučných algoritmov“*

bol splnený následovným spôsobom:

Tretia simulácia ukázala že kombináciou geometrického prístupu[7], evolučnej optimalizácie a vstupných dát obsahujúcich vety jednoduchej materštiny je možné dokonvergovať k schémam - majúcim podobu regulárnych výrazov - ktoré sú schopné vygenerovať syntakticky a sémanticky správne vety, ktoré neboli obsiahnuté vo vstupných dátach.

Štvrtý cieľ: *„Vývoj a zverejnenie evolučného algoritmu schopného vyindukovať gramatické pravidlá z čistého textu“*

bol splnený následovným spôsobom:

Zdrojový kód tretieho programu je zverejnený na adrese http://wizzion.com/thesis/simulation3/EGI.tgz pod licenciou mrGPL.

6 Publikácie autora

El Ghali, A., Hromada, D., and El Ghali, K. (2012). Enrichir et raisonner sur des espaces sémantiques pour l’attribution de mots-clés. In Actes de l’atelier de clôture du huitième défi fouille de texte (DEFT), pages 81–94, Grenoble, France.

Hromada, D. D. (2010a). Quantitative intercultural comparison by means of parallel pageranking of diverse national wikipedias. In 10th International Conference on the Statistical Analysis of Textual Data-JADT 2010, pages 643–651, Rome, Italy. Edizioni Universitarie di Lettere Economia Diritto.

7 Považujeme za vhodné zmieniť sa o tom, že naša roky trvajúca snaha evolučne indukovať samotné regulárne výrazy a riešiť tak problém gramatickej indukcie na čisto symbolickej, negeometrickej, úrovni nebola ani zďaleka taká úspešná ako hybridné riešenie premosťujúce subsymbolické (geometrické) genotypy so symbolickými fenotypmi.

Hromada, D. D. (2010b). smiled : Sourire naturel et sourire artificiel. de l'utilisation d'opencv pour le tracking, la reconnaissaince des expressions faciales et la détection du sourire. Master's thesis, Ecole Pratique des Hautes Etudes, Paris, France.

Hromada, D. D. (2011a). The Central Problem of Roboethics: from Definition towards Solution. MV-Wissenschaft.

Hromada, D. D. (2011b). Initial experiments with multilingual extraction of rhetoric figures by means of perl-compatible regular expressions. In RANLP Student Research Workshop, pages 85–90, Borovec, Bulgaria.

Hromada, D. D. (2013a). Parallel democracy model and its first implementations in the cyberspace. Teoria politica, 3:165–180.

Hromada, D. D. (2013b). Random projection and geometrization of string distance metrics. In RANLP Student Research Workshop, pages 79–85, Borovec, Bulgaria.

Hromada, D. D. (2014a). Comparative study concerning the role of surface morphological features in the induction of part-of-speech categories. In Text, Speech and Dialogue, pages 46–52, Brno, Czech Republic. Springer.

Hromada, D. D. (2014b). Conditions for cognitive plausibility of computational models of category induction. In Information Processing and Management of Uncertainty in Knowledge-Based Systems, pages 93–105, Montpellier, France. Springer.

Hromada, D. D. (2014c). Empiric introduction to light stochastic binarization. In Text, Speech and Dialogue, pages 37–45, Brno, Czech Republic. Springer.

Hromada, D. D. (2014d). Thesis for rigorous examination. Defended on 4.11.2014 at Slovak University of Technology in Bratislava.

Hromada, D. D. (2015a). Conceptual Foundations : Intramental Evolution & Ontogeny of Toddlerese. Propedeutica Didactica. in print. Supplementary material for PhD. dissertation.

Hromada, D. D. (2015b). Genetic optimization of semantic prototypes for multiclass document categorization. In Proceedings of Elitech 2015 conference, Bratislava, Slovak Republic. Slovak University of Technology. Awarded "best paper" prize in "Applied Informatics" track.

Hromada, D. D. (2016a). Narrative fostering of morality in artificial agents: Constructivism, machine learning and story-telling. In L'esprit au-delà du droit: Pour un dialogue entre les sciences cognitives et le droit. Mare et Martin.

Hromada, D. D. (2016b). Reproducible identification of pragmatic universalia in childes transcripts. In Proceedings of 13th International Conference on Statistical Analysis of Textual Data, pages 541–550. Universite Nice Sophia-Antipolis, France.

Hromada, D. D. and Gaudiello, I. (2014). Introduction to moral induction model and its deployment in artificial agents. In Sociable Robots and the Future of Social Relations, pages 209–216. IOS Press.

Hromada, D. D., Tijus, C., Poitrenaud, S., and Nadel, J. (2010). Zygomatic smile detection: The semi-supervised haar training of a fast and frugal system: A gift to opencv community. In Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), pages 241–245, Hanoi, Vietnam. IEEE.

7 Zoznam použitej literatúry

Aycinena, M., Kochenderfer, M. J., and Mulford, D. C. (2003). An evolutionary approach to natural language grammar induction. Final Paper Stanford CS224N June.

Blackmore, S. (2000). The meme machine. Oxford University Press.

Brown, R. (1958). Words and things.

Brown, R. (1973). A first language: The early stages. Harvard U. Press.

Campbell, D. T. (1974). An essay on evolutionary epistemology. The philosophy of Karl Popper, pages 413–463.

Clark, A. (2010). Distributional learning of some context-free languages with a minimally adequate teacher. In Grammatical Inference: Theoretical Results and Applications, pages 24–37. Springer.

Cohen, T., Schvaneveldt, R., and Widdows, D. (2010). Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of Biomedical Informatics, 43(2):240–256.

Dawkins, R. (1976). The selfish gene. Oxford University Press Oxford.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48(1):71–99.

Edelman, G. M. (1987). Neural Darwinism: The theory of neuronal group selection. Basic Books. Fillmore, C. J., Kay, P., and O'connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. Language, pages 501–538.

Fogel, D. B. (1995). Phenotypes, genotypes, and operators in evolutionary computation. In Evolutionary Computation, 1995., IEEE International Conference on, volume 1, page 193. IEEE.

Gärdenfors, P. (2004). Conceptual spaces: The geometry of thought. MIT press.

Goldberg, D. E. (1990). Genetic algorithms in search, optimization & machine learning. Addison-Wesley.

Chomsky, N. (1957). Syntactic structures. Mouton.

Jiménez López, M. D. et al. (2000). Gramar systems: a formal language-theoretic framework for linguistics and cultural evolution.

Kelemen, J. (2004). Miracles, colonies, and emergence. In Formal Languages and Applications, pages 323–333. Springer.

Labov, W. and Labov, T. (1978). The phonetics of cat and mama. Language, pages 816–852.

Lakoff, G. (1990). Women, fire, and dangerous things: What categories reveal about the mind. Cambridge Univ Press.

Lorenz, K. (1973). Die Rückseite des Spiegels: Versuch einer Naturgeschichte menschlichen Erkennens. R. Piper.

MacWhinney, B. and Snow, C. (1991). Childes manual.

Piaget, J. (1974). Introduction à l'épistémologie génétique. Paris, PUF.

Rechenberg, I. (1971). Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution. Dr.-Ing. PhD thesis, Thesis, Technical University of Berlin, Department of Process Engineering.

Rosch, E. (1999). Principles of categorization. Concepts: core readings, pages 189–206.

Rosch, E. and Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive psychology, 7(4):573–605.

Sahlgren, M. (2005). An introduction to random indexing. In Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering, TKE, volume 5.

Salakhutdinov, R. and Hinton, G. (2009). Semantic hashing. International Journal of Approximate Reasoning, 50(7):969–978.

de Saussure, F. (1916). Cours de la linguistique générale.

Sekaj, I. (2004). Robust parallel genetic algorithms with reinitialisation. In Parallel Problem Solving from Nature-PPSN VIII, pages 411–419. Springer.

Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2005). Unsupervised learning of natural languages. Proceedings of the National Academy of Sciences of the United States of America, 102(33):11629-11634.

Slobin, D. I. (1973). Cognitive prerequisites for the development of grammar. Studies of child language development, 1:75–208.

Smith, T. C. and Witten, I. H. (1995). A genetic algorithm for the induction of natural language grammars. In Proc. of IJCAI-95 Workshop on New Approaches to Learning for Natural Language Processing, pages 17–24.

Turing, A. M. (1939). Systems of logic based on ordinals. Proceedings of the London Mathematical Society, 2(1):161–228.

Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. Nature, 171(4356):737–738.

Wall, L. (1994). The perl programming language.

Wolff, J. G. (1988). Learning syntax and meanings through optimization and distributional analysis. Categories and processes in language acquisition.