Evolučné modelovanie ontogenézy rečových kategórií

4 simulácie

Daniel Devatman Hromada[12]
daniel@udk-berlin.de

[1]Slovak University of Technology
Faculty of Electronic Engeneering and Informatics
Department of Robotics and Cybernetics

[2]Université Paris 8
École Doctoralle Cognition, Langage, Interaction
Laboratoire Cognition Humaine et Artificielle

3.6.2016

Table of Contents

PhD. pod dvojitým vedením

CONVENTION INTERNATIONALE DE COTUTELLE DE THÈSE

Vu l'arrêté du 6 janvier 2005 de du ministre de l'éducation nationale, de l'enseignement supérieur et de la recherche de République Française et la § 54 de loi 131/2002 de République Slovaque

Université de Paris 8	Université Technique de Slovaquie
2, rue de la Liberté 93526 Saint–Denis Cédex	Vazovova 5 812 43 Bratislava 1
représenté par son Président **professeur Pascal Binczak**	representé par le Doyen de la Faculté d'Ingenierie Électrique et d'Informatique **doc. RNDr. Gabriel Juhás, PhD.**

Konceptuálne Základy

Takmer 300-stranový *elaborát* usilujúci sa o syntézu troch vedeckých paradigiem:

1. univerzálny darwinizmus (36 strán)
2. vývojová psycholingvistika (50 strán)
3. komputačná lingvistika (63 strán)

Obsahuje taktiež 38 stranový súhrn kvalitatívnych pozorovaní jedného ľudského toddlera (0-30 mesiacov) a 27 strán kvantitatívnych analýz vyextrahovaných z korpusu Child Language Data Exchange System (CHILDES).

Základné Tézy

1. "Myseľ sa vyvíja" (mind evolves)
2. "Učenie je formou evolúcie"
3. "Učenie možno úspešne simulovať pomocou evolučných výpočtov"
4. "Učenie prirodzených jazykov možno úspešne simulovať pomocou E.V."
5. "Ontogenézu detskej reči možno úspešne simulovať pomocou E.V."

Teória Intramentálnej Evolúcie

Základný postulát

Vývoj individuálnej mysle možno interpretovať - resp. dokonca simulovať - ako proces replikácie, variácie a selekcie v mysli obsiahnutých a informáciu nesúcich kognitívnych štruktúr.

O niečo podobné usilovala už aj Piagetova *genetická epistemológia*, T.I.E. však hovorí aj o simulácii či dokonca emulácii...

Simulácie mojej dizertácie sú snahou o poskytnutie určitého dôkazu *ex computatione* platnosti tejto teórie.

Table of Contents

Voyničov rukopis

Enigma

240 strán textu napísaných v neznámom písme (a možno aj v neznámom jazyku) sprevádzaných ilustráciami s motívami botaniky, zdravovedy, astrológie atď.

Nultá simulácia

1. môj prvý vlastný evolučný algoritmus
2. genóm každého jedinca má dĺžku 19 znakov a udáva možný prepis jedného symbolu v rukopise na jednu z možných foném výsledného jazyka (napr. slovanské jazyky 38 znakov)
3. sústreďuje sa na prepis jednej časti rukopisu, tzv. "kalendár" na zoznamy krstných mien
4. prepisy sú najúspešnejšie keď slovníky obsahujú ženské mená písané zprava doľava
5. hebrejské a slovanské zdrobnelé ženské mená...

Spoločné črty simulácií 1-3

Všetky tri simulácie

1. sa usilujú o riešenie problémov strojového učenia
2. používajú texty písané v hovorovej angličtine ako vstupné dáta
3. charakterizujú slová v týchto textoch pomocou ich určitých čŕt: tieto črty sú následne využité v premietnutí textu do vektorových priestorov
4. principiálne operujú v relatívne nízkorozmerných binárnych (Hammingových) priestoroch
5. uskutočňujú evolučné vyhladávanie optimálnych riešení
6. v najvnútornejšom cykle vyhodnocovania účelovej funkcie vždy dochádza k meraniu Hammingových vzdialeností

Viactriedna sémantická klasifikácia textov

- Elitech 2015, aplikovaná informatika (ocenenie)
- Korpus: 20 newsgroups (18845 textov z 20tich usenetových kategórií)
- 11314 textov: trénovacie dáta; 7543 textov: testovacie dáta
- frekvencie výskytov jednotlivých slov v jednotlivých textoch udávajú črty pomocou ktorých text geometrizujeme

Základná idea

Vo vektorovom priestore vyhľadávame také body ktoré sú čo najbližšie k vek.rep. objektov určitej kategórie a čo najďalej od vek.rep. objektov iných kategórií.

Teória Prototypov

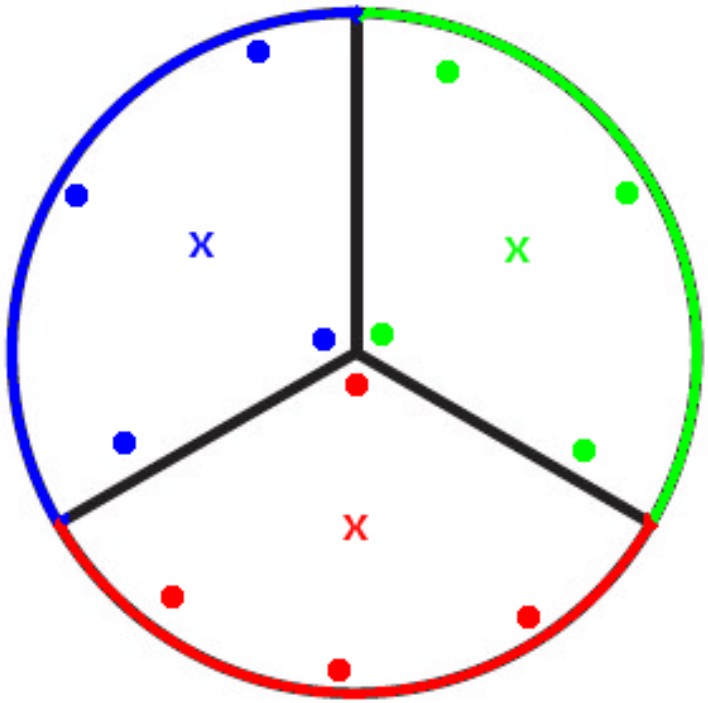
Items rated more prototypical of the category were more closely related to other members of the category and less closely related to members of other categories than were items rated less prototypical of a category (Rosch a Mervis, 1975)

Fitness funkcia:

$$F_{CP}(P_K) = \sum_{t \in C_K} F_{hd}(h_t, P_K) - \sum_{f \not\in C_K} F_{hd}(h_f, P_K) \qquad (1)$$

(P_K kandidát na prototyp K-tej triedy; h_t vektorová reprezentácia objektu tiež náležiaceho do K; h_f vektorová reprezentácia objektu ktorý do K nepatrí; F_{hd} Hammingová vzdialenosť)

Problém lineárnej oddeliteľnosti...



...možno nieje pre klasifikačné modely založené na Teórii Prototypov až takým pálčivým problémom !

Učenie slovných druhov

Problémy ako *part-of-speech (POS) induction* a *POS tagging* sú jedny z najlepšie rozpracovaných problémov výpočtovej lingvistiky.

O užitočnosti slovných druhov

1. Ak človek dokáže rozpoznať že neznáme slovo W_X patrí do kategórie K, dokáže mu ľahšie priradiť význam.
2. Bez slovných druhov nieto gramatík.

Druhá simulácia:

1. sekcia Brown / Eve korpusu CHILDES
2. prepisy POS tagy manuálne opravené ľudskými anotátormi
3. trénovací korpus (972 slovných typov) : Eve pred dosiahnutím dvoch rokov veku; testovací korpus (934 slovných typov): Eve vo veku 2 - $2.\frac{1}{2}$ roka
4. 449 slovných typov sa vyskytuje **iba** v testovacom korpuse

Metóda

Iba tri jednoduché črty sú použité na priemet slovo X do vektorového priestoru: prípona slova X, prípona slova napravo od X a prípona slova naľavo od X.

Operačný princíp A

Pay attention to the ends of words. (Slobin, 1973)

Po geometrizácii všetkých tokenov následne vyhľadávame prototypy jednotlivých tvaroslovných tried pomocou účelovej funkcie

$$F_{object}(\vec{i}, \vec{o}) = \underset{p_x \neq p_T \wedge Hd(\vec{o},\vec{p_x}) <= Hd(\vec{o},\vec{p_T}) \Longrightarrow p_x \hookrightarrow P_F}{|P_F|} \quad (2)$$

t.j. penalizujeme za každý nesprávny prototyp p_X ktorý je k objektu $\vec{o}$ bližšie ako správny (p_T).

To čo vyhľadávame sú optimálne **konštelácie** prototypov.

Zopár výsledkov

Table 5: MSVM2 training corpus confusion matrix.

	ACT	SUB	PROP	REL	REF
ACT	266	54	0	0	1
SUB	55	495	4	0	0
PROP	21	66	18	0	0
REL	20	12	1	2	0
REF	15	47	3	0	0

Table 6: MSVM2 testing corpus confusion matrix.

	ACT	SUB	PROP	REL	REF
ACT	271	38	4	0	1
SUB	55	450	8	1	0
PROP	21	62	15	0	0
REL	20	6	3	0	0
REF	20	38	5	0	0

Table 7: Training corpus confusion matrix produced by FITTEST(GA_{MERGE1}).

	ACT	SUB	PROP	REL	REF
ACT	278	28	4	4	7
SUB	56	427	34	18	19
PRO	19	39	43	3	1
REL	15	5	1	11	3
REF	9	35	7	1	13

Table 8: Testing corpus confusion matrix produced by FITTEST(GA_{MERGE1}).

	ACT	SUB	PROP	REL	REF
ACT	269	21	9	6	9
SUB	62	371	41	25	15
PRO	16	35	40	3	4
REL	15	3	4	5	2
REF	11	26	8	4	14

Výsledky čo prekvapili...

A subsequent inspection of false positives turns out to be quite instructive. Hence, the token **"building", present in the utterance "what are you building here?" on line 5417 of eve05.cha transcript is clearly not a noun, as CHILDES annotators and correctors supposed**, but rather a participle - and hence an instance belonging to ACTION class, as correctly predicted by $FITTEST(GA_{MERGE1})$. **Idem for "hit" present in the utterance "did you hit your head?"** present on line 4145 of eve01.cha transcript: the token is clearly not a noun, as postulated by CHILDES annotators, but, as predicted, a verb and hence member of ACTION class. And one can continue: **the token "matter" annotated on lines 2152 and 5688 of CHILDES corpus as a verb is clearly not a verb but a noun - and hence a member of a class SUBSTANCE - because it twice occurs in the utterance "what's the matter?**. And **in spite of the fact that CHILDES labels the token "numbers" as a verb, it is definitely not a verb when it occurs in the utterance "the numbers are going around too"** (eve15.cha, line 6276). Et caetera et caetera.

Indukcia | Inferencia Gramatiky

Definícia problému

Máme množinu M viet jazyka J. Cieľom IG je vydestilovať z M poznatky (resp. model, pravidlá, schémy, vzory atď.) ktoré nám následne umožnia vygenerovať aj také vety jazyka J ktoré neboli v M.

Kameň úrazu

Prílišné zovšeobecnenie (over-generalisation resp. over-regularisation): napr. keď dvojročné dieťa začne hovoriť *goed* namiesto *went*.

Cieľom IG je nájsť také systémy pravidiel ktoré niesú ani príliš špecifické: $(1 \rightarrow < corpus >)$, ale ani príliš všeobecné:

$$1 \rightarrow 2*$$

$$2 \rightarrow a|b|c \ldots Z$$

Klenbový svorník

Problém prílišného zovšeobecnenia možno vyriešiť tak, že nastavíme evolučný proces spôsobom ktorý bude penalizovať príliš všeobecné riešenia. Schopnosť evolúcie zbaviť sa toho čo je nepotrebné sa postará o zvyšok.

$$Fitness_1(N_X) = \frac{Y_X * Y_X}{E_X}$$

kde Y_X je počet viet korpusu matchnutých fenotypickým prejavom $N-$schémy X a E_X je teoreticky maximálna možná daná extenzia

$$E_X = \prod_{k=1}^{N} I_{H_k}$$

získaná ako multiplikatívny produkt extenzií kategórií ktoré su v N_X kódované.

Od teórie k praxi

Theoria

$\Delta-$rozmerné vektorové priestory, G-kategórie, Hammingové sféry, H-kategórie, Syntagmatické a paradigmatické kategórie, N-schémy...

Praxis

Prepis vektorov ktoré popisujú konštelácie oblastí v hammingových priestorov na staré dobré PERLovské regulárne výrazy.

Syntagma	H_1		H_2		H_3		H_4		H
	Center	Radius	Center	R	Center	R	Center	R	Cente
	BABC	17	0F20	5	5FF0	7	C124	3	7723

^(this |that|it)(is)(not)(a |the)(dog |duck)$

Prvé výsledky

c.f. Appendix 1

Pár otázok

Možno pomocou evolučných algoritmov realizovať strojové učenie ?

ÁNO: V prípade že ústrednou črtou strojového učenia je schopnosť zovšeobecniť poznatky obsiahnuté v trénovacích dátach.

Môžu byť evolučné algoritmy užitočné na riešenie problémov výpočtovej lingvistiky ?

ÁNO: Ale len za predpokladu vhodne zvolenej účelovej funkcie a reprezentácie jednotlivých riešení.
Odporučenie: kombinácia subsymbolických (napr. geometrických) a symbolických úrovní reprezentácie sa ukazuje ako užitočná.

Výhody evolučného prístupu v porovnaní s konekcionistickými riešeniami?

Konekcionisti modelujú štrukturálne vlastnosti kognitívnych systémov. Ale možnosť definovať fitness funkciu umožňuje

Ďakujem za pozornosť.