Predbežná prehľadová kapitola

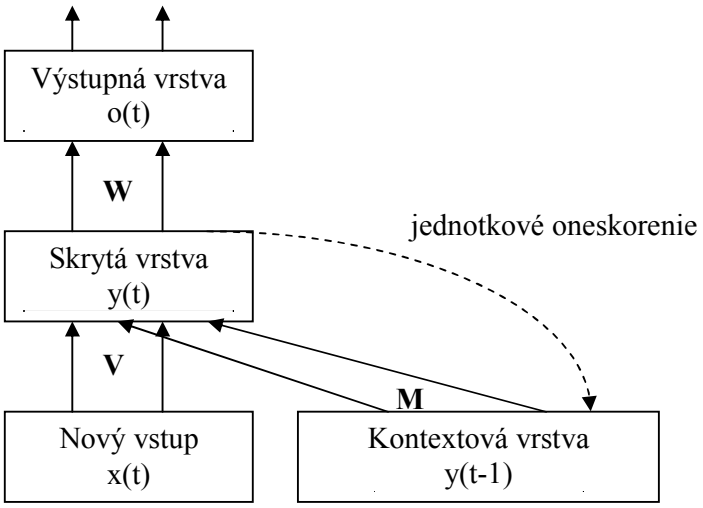
Modelovanie porozumeniu jazyka je zložitý problém pre neurónové siete (konekcionizmus) a preto súčasné modely sa zaoberajú zjednodušenými, prípadne umelými gramatikami. V týchto systémoch, ktoré simulujú osvojovanie jazyka sú častým problémom gramatické vzťahy. Predstavujú najabstraktnejší aspekt jazyka. Sémantika je už spájaná s významom slov, syntax je nezávislá na tom čo ktoré slovo znamená. Preto v teóriach jazyka sú za fundamentálny aspekt považované gramatické vzťahy.
Psycholingvisti sa domnievajú, že existuje časť (respektíve časti) mozgu, ktorá sa zaoberá čisto spracovávaním syntaxe a je nezávislá na sémantike. Toto je podporené dátami získanými pomcou fRMI. Preto je snaha tento komponent modelovať. Na túto úlohu sa ukázal byť vhodný model neurónovej siete SRN. Niektoré práce využívajúce SRN popíšem neskôr.

Model SRN (simple recurrent network)

Tento model rekuremtnej siete sa využíva pri spracovávaní vstupov, ktoré pozostávajú zo sakvencií rôznej dĺžky. Napríklad veta, sekvencie sú slová. Sieť zohľadňuje nielen momentálny vstup, ale aj predchádzajúce. To isté slovo môže mať rôzne syntaktické role (napr. podstatné mená môžu vystupovať ako objekt aj subjekt). Preto nestačí len namapovať slová na vetné členy. SRN spracúva vstup na základe časového kontextu. Toto dovoľuje zohľadniť informácie z predchádzajúcich prvkov postupnosti.

Sieť sa skladá z vrstiev neurónov. Vrstvy sú pospájané ako na obrázku 1, t.j. každý neurón so vstupnej vrstvy je spojený s každým so skrytej vrstvy atď. Dvojité šípky zobrazujú spojenie vrstiev. Ku každému spojeniu je priradená váha, nejaké reálne číslo. Dostávame tak váhové vektory M, V a W.

Architektúra SRN:



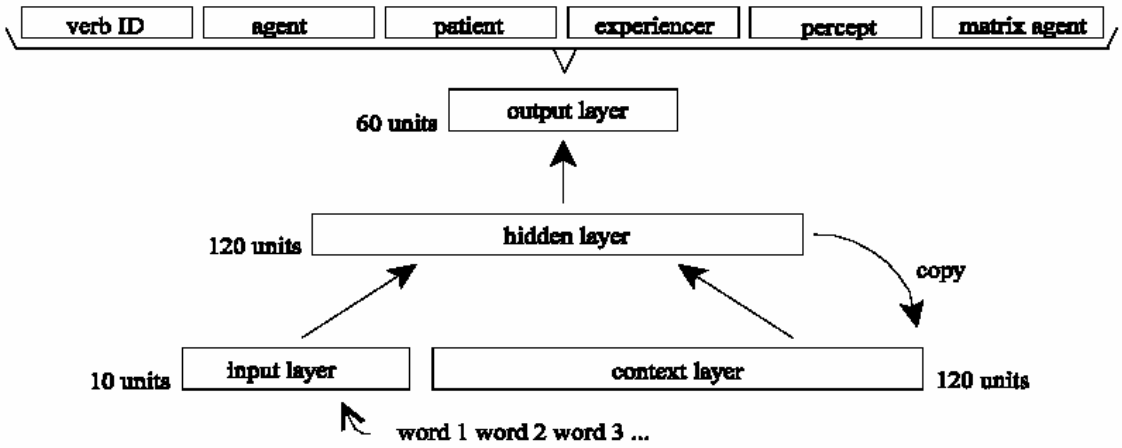
Obrázok 1: Architektúra Elmanovej SRN.

V čase t je na vstupe vektor $\bar{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, ..., x_I^{(t)})$, skutočný výstup je vektor $\bar{o}^{(t)}$ a požadovaný výstup je $\bar{d}^{(t)} = (d_1^{(t)}, d_2^{(t)}, ..., d_K^{(t)})$. Váhy sa upravujú v trénovacích

cykloch(epochách) pomocou metódy známej ako rekurentné učenie v reálnom čase. Rovnice tohto postupu pre SRN z obrázka 1 možno nájsť v [1].

Simulácia pomocou SRN na umelom jazyku podobnému angličtine (podľa [4])

Bolo skúmané či sieť dokáže vytvoriť abstraktné vzťahy korešpondujúce z objektom a subjektom v jednoduchom jazyku anglického typu. Sieť nemala žiadne "vrodené" vedomosti o jazyku. Sieť dostávala na vstup sekvencie vzoriek, ktoré reprezentovali vety generované gramatikou. Po každej vete nasledovala špeciálna vzorka "reset".
Architektúra použitej SRN:



Slovník pozostával z 56 slov, ktoré boli reprezentované ako 10-bitové vzorky. Z toho bolo 25 slovies a 25 podstatných mien, 6 zostávajúcich prislúchalo ostatným slovným druhom. Všetky slová boli legálne slová anglického jazyka. Sieti bolo predložených 50 000 viet na natrénovanie.

Veta bola určená správne, ak každému slovu bola priradená jeho správna syntaktická rola. Priemerná úspešnosť bola 76%, avšak pri jednoduchších vetách bola až 91-97%.

Simulácie Lupyana a Christiansena

Pre slovenčinu nie sú známe žiadne podobné experimenty, ale zaujímavé simulácie urobili Lupyan a Christiansen [3]. Porovnávali schopnosti neurónových sietí naučiť sa syntax umelých jazykov. Ako pomôcka slúžilo buď striktné poradie vetných členov (SWO-strict word order, jazyky ako angličtina), alebo pádové koncovky slov (napr. slovanské jazyky). Ak boli prítomné koncovky tak poradie slov bolo väčšinou voľné (FWO-free word order). Toto vychádza z reality, lebo svetové jazyky môžeme vo všeobecnosti rozdeliť do týchto dvoch skupín.

V troch simuláciach skúmali, či sieti stačia pádové koncovky a, resp. alebo poradie vetných členov na osvojenie si syntaxe. V prvej simulácii uvažovali 6 možností SWO jazykov (kombinácie podmetu, prísudku a predmetu), a FWO jazyk. To všetko s alebo bez pádových koncoviek, čiže spolu 14 umelých gramatík. Cieľ bolo zistiť ako koreluje cshopnosť siete naučiť sa daný typjazyka s výskytom daného typu v skutočnosti.

Siete trénované na jazykoch s koncovkami mali všetky 100% úspešnosť. Keď pád nebol prítomný tak výkon siete hrubo zodpovedá frekvencii výskytu jazyka. Iba poradie SVO(subject-verb-object) a VSO(verb-subject-object) mali úspešnosť skoro dokonalú, 99%. A práve väčšina bezpádových jazykov je typu SVO, alebo VSO. Najnižšia úspešnosť bola očakávateľne pri FWO, iba 65%. V realite však všetky FWO jazyky majú skloňovanie.

V prirodzených jazykoch, nie sú koncovky úplne deterministické. V slovenčine existujú slová, ktoré sa neskloňujú. Naviac koncovky ovplyvňujú výslovnosť a preto dochádza k nejednoznačnostiam. V druhej simulácii tak vytvorili 5 gramatík líšiacich sa hustotou výskytu označenia pádov. Mezdi nimi bola aj umelá gramatika podobná poľštine. Tá mala pri 75% prítomnosti koncoviek 90% úspešnosť. Pri 100% prítomnosti pádových značiek boli opäť všetky siete 100%-ne úspešné.

Na simulácie boli použité SRN. Slová boli náhodne generované 20-bitové vektory. Pádové koncovky boli 4-bitové vektory pridávané ku vstupným slovám. Sieť mala 7 výstupných neurónov reprezentujúcich 7 gramatických rolí: podmet, priamy predmet, nepriamy predmet, podstatné meno v genitíve, sloveso a koniec vety. V skrytej, aj kontextovej vrstve bolo 30 neurónov. Slovník mal 300 podstatných mien a 100 slovies. Siete sa trénovali v 100 000 epochách.

Výsledky potvrdili, že ľahkosť naučenia sa syntaxe jazyka môže byť hlavný faktor pri frekvencii výskytu danej triedy jazykov. Jazyky ktoré sú ľahko naučiteľné ľuďmi prežívajú. Tie ktoré sa učia ťažko zanikajú, alebo nikdy nevzniknú. Simulácie dokazujú, že k naučeniu syntaktických vzťahov stačí spoľahlivá pomôcka: Pevné poradie slov, alebo pádové koncovky.

Slovenčina patrí do skupiny jazykov s voľným poradím slov, ale má pádové koncovky. Preto sa domnievam, že sieť typu SRN sa bude schopná naučiť predikovať vetné členy na jednoduchom vetníku.

Literatúra:

[1] Beňušková L. Umelé neuronové siete . Kapitola z Návrat P. et al. *Umelá inteligencia.* STU: Bratislava, 2002, str. 161-189, rozšírená o BPTT, RTRL, IFS a SOM.
[2] Palmer-Brown, D., Tepper, J.A. a Powell, H.M. Connectionist natural language parsing, Trends in Cogn. Sci., 2002
[3] Lupyan, G.a Christiansen, M.H.: Case, Word Order, and Language Learnability: Insights from Connectionist Modeling
[4] Morris, W.C., Cottrell, G.W. a Elnam, J., A Connectionist Simulation of the Empirical Acquisition of Gramatical Relations