Rozbor chyb v automatickém stochastickém parsingu pomocí MST parseru

V této zprávě přinášíme dva nezávislé rozbory chyb ve výsledcích stochastického parsingu s pomocí MST parseru Ryana McDonalda: plně automaticky získané statistiky chyb testované proti pečlivě manuálně anotovaným datům a frekvence některých typů chyb, poloautomaticky vyhledaných ve stomiliónovém, syntakticky anotovaném korpusu SYN2005.

1. Plně automaticky získané statistiky chyb MST parseru

Pro spolehlivé a důkladné testování chyb parseru jsme vyvinuli metodu, která umožňuje využít veškerá manuálně anotovaná data zároveň k trénování i k testování (test se tedy neprovádí jen na cca 20% dat, je spolehlivější). Všechna manuálně syntakticky označkovaná data PDT (trénovací i testovací) byla rozdělena na pět testovacích množin po 20 % objemu, zbývající data byla vždy použita jako trénovací. Parser se tak vždy testoval na „neznámých" datech, ale výsledkem bylo celých cca 1,5 miliónu slov stochasticky označkovaných, jež bylo možné srovnat s „gold" daty a zjistit tak úspěšnost.

Uvádíme dvě tabulky a jeden graf: tabulku chybovosti určení řídícího členu v závislosti na slovním druhu slova (dle morf. anotace PDT, vč. interpunkce a neznámých slov), tabulku chybovosti určení řídícího členu v závislosti na syntaktické funkci větného členu a jeho příslušnosti do koordinace, apozice či parenteze a graf chybovosti v závislosti na vzdálenosti a pozici řídícího větného členu.

1.1 Tabulka chybovosti určení řídícího členu podle slovního druhu

POS	podíl chyb	frekvence
N	**13.31**	303456
A	**6.27**	117745
P	**6.45**	65474
C	**17.92**	31955
V	**18.72**	122152
D	**17.95**	52607
R	**20.45**	97034
J	**28.23**	56418
T	**25.06**	5318
I	**40.18**	74
Z	**20.26**	147752
X	**92.31**	8

V prvním sloupci tabulky jsou slovní druhy (první znak z morfologické značky u manuálního značkování), ve druhém sloupci procentuální podíl chybných určení řídícího členu ze všech výskytů (zeleně je zvýrazněna chybovost pod 10 %, červeně chybovost nad 50 %), ve třetím sloupci je frekvence slovního druhu přepočítaná na 1 milión slov.

Parser zvládá velmi dobře určovat řídící členy u adjektiv a u zájmen. Výrazně nadprůměrnou chybovost má parser jednak u málo frekventovaných slovních druhů (X, I, T), jednak u předložek, spojek a interpunkce.

1.2 Tabulka chybovosti určení řídícího členu podle synt. funkce

Ve druhé tabulce je chybovost rozdělena podle (správné) syntaktické funkce a případné příslušnosti slova do koordinační nebo apoziční skupiny, popř. parenteze. Pro každou tuto kombinaci (např. podmět v koordinaci: Sb + Co) je uveden procentuální podíl chybných určení řídícího členu a frekvence v rozšířených testovacích datech přepočítaná na 1 milión slov.

AFUN	-	-	**Co**	Co	**Ap**	Ap	**Pa**	Pa
Pred	**10.65**	37073	**29.12**	21394	**53.76**	300	**79.05**	984
Sb	**12.35**	62643	**23.26**	7891	**49.64**	2395	**75.00**	5
Pnom	**11.32**	13298	**24.62**	1093	**57.54**	119		
Atr	**10.49**	259032	**17.65**	24774	**49.18**	1998	**51.13**	557
Obj	**8.47**	70687	**21.86**	10243	**47.97**	1394	**33.33**	15
Adv	**8.51**	95107	**16.27**	7405	**37.11**	1057	**26.16**	528
Atv	**63.25**	1581	**26.83**	136	**56.00**	16	**78.95**	12

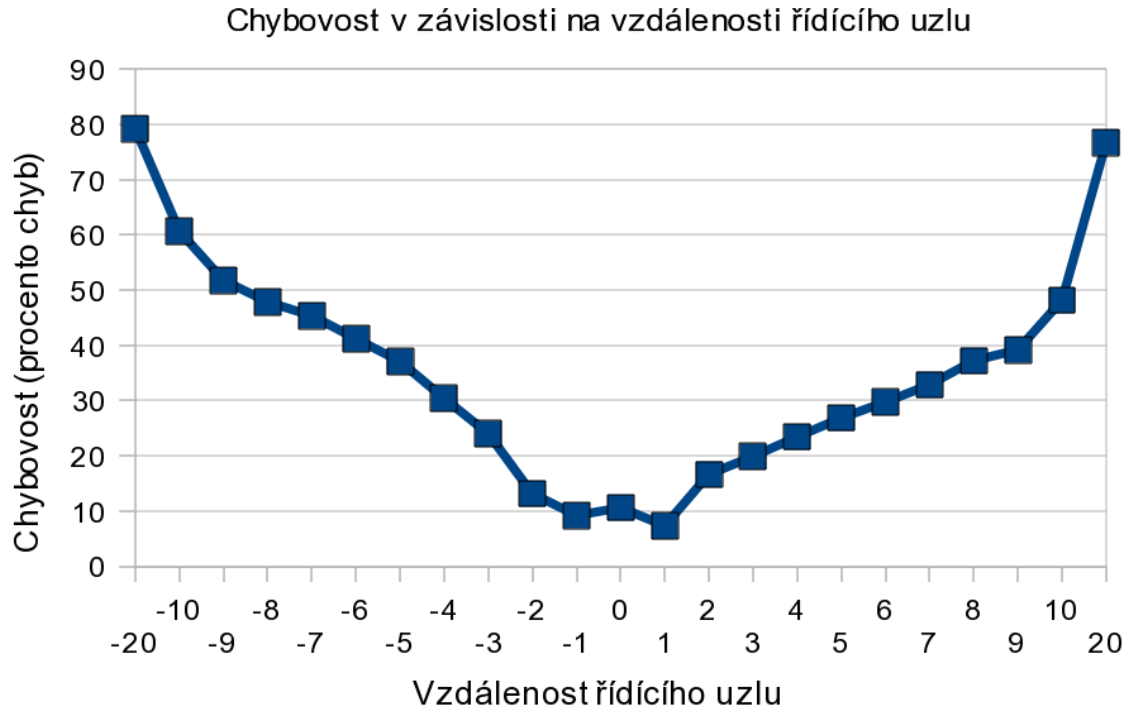
AFUN	-	-	Co	Co	Ap	Ap	Pa	Pa
AtvV	**26.54**	606	**17.74**	41	**50.00**	2	**87.50**	5
ExD	**32.93**	18307	**29.94**	11553	**30.33**	1639	**52.64**	2989
AuxV	**8.31**	11992	**54.55**	7				
AuxP	**20.54**	99110					**57.14**	4
AuxC	**20.53**	17012	**91.67**	7			**50.00**	1
Coord	**36.25**	35935	**62.55**	3303	**62.63**	1110	**70.78**	910
Aux[GKX]	**14.61**	127484	**91.67**	7			**21.43**	9
Aux[OYZ]	**22.62**	23765	**65.22**	15			**48.68**	50
Aux[RT]	**4.12**	14687						
Apos	**55.81**	4478	**69.28**	441	**68.87**	100	**82.14**	18
AtrAdv	**5.99**	1477	**11.67**	170	**38.10**	13	**12.50**	5
AtrAtr	**10.17**	510	**6.45**	20				
AdvAtr	**5.03**	343	**13.16**	25	**0.00**	3		
AtrObj	**23.64**	36	**12.50**	5				
ObjAtr	**16.67**	27	**0.00**	1				

Opět je z tabulky zřejmé, že případy, s nimiž se parser setká v trénovacích datech jen málo, jsou značkovány velmi špatně. Zde to platí především pro parentezi, jež se vyskytuje relativně zřídka a její určení navíc z velké části záleží na sémantice: automaticky je tak velmi obtížné ji určit.

1.3 Graf chybovosti určení řídícího uzlu v závislosti na vzdálenosti a pozici řídícího uzlu

Poslední analyzovaný typ chyby zobrazujeme graficky. V grafu je uvedena „správná“ (tedy manuálně určená) vzdálenost řídícího uzlu od zkoumaného větného členu: záporná čísla znamenají, že řídící člen je vlevo od zkoumaného slova (před ním), kladná čísla znamenají, že řídící člen je vpravo od daného slova. „Vzdálenost“ nula označuje řídící slovo ve větě (obvykle predikát v hlavní větě). -20, -10, 10 a 20 označují součty vyšších hodnot (deset a více, dvacet a více atd.).

Jak vidíme, se vzrůstající vzdáleností řídícího členu roste i chybovost. MST parser s větší spolehlivostí určuje závislost „vpravo“ (tj. na jednom z následujících slov) než „vlevo“, i když se jejich frekvence nijak výrazně neliší.



2. Poloautomatické vyhledávání chyb v syntakticky anotovaném korpusu SYN2005

Během grantového projektu byl MST parserem syntakticky anotován korpus SYN2005. Výsledky parsingu byly manuálně analyzovány, zjišťovaly se opakující se typy chyb, které by bylo možné automaticky identifikovat (případně i automaticky opravit). Byl vytvořen program pro vyhledávání chyb, schopný najít cca 50 typů chyb: celkem ale identifikoval jen asi 1/10 všech předpokládaných chyb, které se v korpusu při chybovosti okolo 20 % musely vyskytnout.

Automaticky identifikované chybné struktury jsou představeny v tabulce. Tato analýza byla východiskem pro opravný program popsaný ve zprávě **Rule-based_correction_system.pdf**.

V prvním sloupci tabulky jsou typy identifikovaných chybných struktur, v druhém sloupci frekvence v korpusu SYN2005, ve třetím sloupci podíl dané chyby na všech automaticky identifikovaných chybách. Uvádíme pouze struktury s frekvencí 20 000 výskytů a více.

Typ automaticky identifikované chyby	frekvence v SYN2005	podíl %
Pred u slovesa v koordinaci závislého na jiném větném členu	401234	20,1
Nekompatibilní funkce v koordinaci (**Obj**+**Atr**, **Adv**+**Atr**, **Sb**+**Atr**)	299810	15,0
Dva subjekty závislé na témž slovese (ne koordinované)	261888	13,1
AuxT u zájmena závislého na slovese, jež není refl. tantum	241965	12,1
Pred u slovesa závislého na jiném větném členu (ne v koordinaci)	225431	11,3
Obj u substantiva - nominativu	129539	6,5
Adv u osobního zájmena (P[PH7])	98023	4,9
Sb u substantiva - nenominativu (ne však genitiv záporový)	93777	4,7
Chyby valence u předložkových frází	83572	4,2
Atr závislý na slovesu (přes předložku)	64019	3,2
Obj/**Sb** závislé na substantivu (přes předložku)	47366	2,4
Obj u substantiva v akuzativu závislého na slovese (ne "učit", "dozvědět"), kde také visí zájmeno "se" (také chyba značkování)	31806	1,6
Dvě substantiva různých pádů v koordinaci pod předložkou	29198	1,5
Adv závislé na substantivu (přímo)	27293	1,4
Adv závislé na substantivu (přes předložku)	22895	1,1