ILP a český korpus. Rozšířený abstrakt

30. září 1998

Dlouhodobým cílem této části projektu je vývoj systému pro podporu desambiguace českých vět, který bude využívat metod induktivního logického programování(ILP) [9].

ILP bylo úspěšně aplikováno při segmentaci slov [6], v morfologické [7], syntaktické i sémantické analýze [8] stejně jako pro kategorizaci textů [1]. V oblasti morfologie v LAI Ljubljana [4] použili ILP pro generování 1. pádu jednotného čísla a nalezení správného tvaru v daném pádu s průměrnou přesností 91.5 % [4]. Nalezení správných morfologických kategorií (značkování, tagging) je jednou z nejčastějších a z nejúspěšnějších aplikací ILP. V [2] je popsán značkovač pro angličtinu, který dosahuje přesnosti 96.4 %, která je srovnatelná s přesností nejlepších existujících značkovačů. Indukce pravidel v constraint grammar je popsána v [5].

Uvádíme první výsledky při zpracování korpusu českých vět DESAM.

1 ILP a český korpus: řešené úlohy

Cílem první fáze našeho projektu bylo vyzkoušení schopností ILP pro tvorbu podpůrných modulů pro desambiguaci. Kromě rozpoznání konce věty, rozpoznání jmenné fráze a rozpoznání podmětné a přísudkové části věty (viz [13]) jsme se zaměřili především na

- Desambiguace homonym
- Desambiguace na úrovni morfologických kategorií
- Zjemnění morfologických kategorií

Jako *doménová znalost* byly použity jen predikáty testující výskyt určitého tagu. Byly použity systémy Progol [10] a WiM [12]. Úplný popis výsledků lze nalézt v [13].

2 Desambiguace na úrovni lemmatu

Desambiguací na úrovni lemmatu se rozumí jednoznačné přiřazení lemmatu, tj. základního tvaru, takovému slovnímu tvaru, kterému lze z hlediska morfologie přiřadit více lemmat. V závislosti na kontextu je však správné právě jedno lemma.

V korpusu DESAM se vyskytuje mnoho slovních tvarů s více lemmaty. Pro účel experimentů byly vybrány tři slovní tvary

Slovní tvar	lemma	Počet výskytů	Úspěšnost
časem	čas	56	88.89 %
	časem	13	75.00 %
stranou	strana	27	100.00 %
	stranou	16	50.00 %
středu	střed	56	38.46 %
	středa	95	90.00

Průměrná úspěšnost je tedy 73.7

3 Desambiguace na úrovni gramatických kategorií

Konkrétně jsme se zameřili na podčást desambiguace - určení morfologických kategorií podléhajících vzoru `stroj` pomocí systému P-Progol. Pro učící množinu obsahující 130 vět výpočet trval necelé půl minuty. Nedostatkem je poměrně velký počet klauzulí(40). Získaná množina pravidel byla testována na 120 příkladech se slovním tvarem olej a na 60 příkladech s tvarem zdroj. Akceptovala asi 75% správně označkovaných slov a nepřijala přibližně stejně velký počet nesprávně označkovaných.

4 Zjemnění gramatických kategorií

Morfologická kategorizace slovních tvarů prováděná morfologickým analyzátorem Lemma je často příliš hrubá a je zapotřebí ji zjemnit. Za současného stavu se značky poskytované Lemmou nechovají jako názvy gramatických kategorií ve smyslu matematické lingvistiky [11]. Shledali jsme, že v češtině existují slovní tvary x,y s totožnou značkou a dále takový kontext (u,v) - říkejme mu rozlišující kontext -, že uxv je správně utvořená česká věta, tedy tvar x je přijímán kontextem (u,v), zatímco uyv není syntakticky správná věta, kontext (u,v) nepřijímá tvar y. Tvary x,y tudíž nemohou náležet do téže gramatické kategorie.

```
Přinesl < metr > vysoký sloupec novin.
* Přinesl < vzduch > vysoký sloupec novin.

< Má > na peníze právo.
* < Věří > na peníze právo.
```

Kromě těchto případů, které vedou k zavedení zcela nových, "hybridních" kategorií, navíc existují slovní tvary, pro něž nedává Lemma ani všechny ty značky, které by daný tvar mohl mít už ve stávajícím systému, např.

`-- př.: Potkal jsem ji < cestou >.`

Cílem projektu bylo sestrojit predikát, který by umožnil rozpoznávat ony rozlišující kontexty od kontextů, ve kterých se obě daná slova chovají syntakticky totožně, tedy tak, že jejich vzájemná záměna je možná bez porušení syntaktické správnosti věty. K řešení stanoveného cíle byl použit ILP systém Progol. Pro několik vybraných slovních tvarů byly systému předloženy jejich rozlišující kontexty jako pozitivní příklady, ostatní kontexty pak jako příklady negativní. Výpočetní doba potřebná k nalezení ideální teorie byla pro 20 zadaných příkladů (10 pos., 10 neg.), s použitím omezujících podmínek na tvorbu klauzulí, zhruba 5 sekund. Nalezená teorie obsahovala pět klauzulí, z nichž každá měla nejvýše dva literály v těle.

5 Závěr

Základním cílem naší práce je podpora desambiguace českých korpusu. Dosažená průměrná úspěšnost určení lemmatu (73.7 %) je oproti očekávání nižší, především pro nízkou úspěšnost u slova `středu` pro lemma střed. Dá se očekávat, že s ILP systémem je možno dosáhnout mnohem vyšší úspěšnosti při použití bohatší doménové znalosti.

Co se týče desambiguace na úrovni gramatických kategorií, zdá se vhodné tuto metodu skombinovat se statistickou; pokud nebyla nalezena značka jednoznačně, nabídne se uživateli nejdříve ta, která je nejčastější.

Rozšíření doménové znalosti, např. predikáty pro jmenné a slovesné fráze, pro shody v číslech a rodech umožní nepochybně zvýšit úspěšnost.

Srovnáme-li výsledky s výsledky dosaženými buď pomocí n-gramů nebo neuronových sítí, nejsou o mnoho horší, přihlédneme-li k tomu, že v experimentech nebyla použita žádná hlubší lingvistická znalost.

Odkazy

[1] Cohen W.: Text Categorization and Relational Learning. In Proc. Of 13th Int. Conf, on Machine Learning ML'95, 1995.

[2] Cussens J.: Part-of-Speech Tagging using Progol. In Proc. of ILP'97, LNAI 1297, Springer-Verlag 1997.

[3] Dehaspe L., De Raedt L.: Mining Association Rules in Multiple Relations. In Proc. of ILP'97 pp.125-133, LNAI Springer-Verlag

[4] Džeroski S., Erjavec T.: Induction of Slovene Nominal Paradigms. In Proc. of ILP'97, LNAI 1297, Springer-Verlag 1997.

[5] Eineborg, Lindberg: Induction of Constraint Grammar rules using Progol. In Proc. of ILP'98, 1998.

[6] Kazakov D., Manandhar S.: Hybrid approaches to word segmentation. In Proc. of ILP'98, 1998.

[7] Manandhar S, Džeroski S., Erjavec T.: Learning multilingual morphology with CLOG. In Proc. of ILP'98, 1998.

[8] Raymond J. Mooney: Inductive Logic Programming for Natural Language Processing. Inductive Logic Programming: Selected Papers from ILP'96 S. Muggleton (Ed.), pp.3-22, Springer Verlag, Berlin, 1997.

[9] Muggleton S., De Raedt L.: Inductive Logic Programming: Theory And Methods. J. Logic Programming 1994:19,20:629-679.

[10] Muggleton S.: Inverse Entailment And Progol. New Generation Computing Journal, 13:245-286, 1995

[11] Novotný, M.: S algebrou od jazyka ke gramatice a zpět. Academia, Praha 1988.

[12] Popelínský L., Štěpánková O.: WiM : A Study On Top-Down ILP Program(abstract). In: Proceedings of AIT'95 Workshop, Brno, Czech Republic, ISBN 80-214-0673-9(Abstract). FI MU Report FIMU-RS-95-03, 1995

[13] Popelínský L. et al: ILP a český korpus. První zpráva k tomuto grantu, 1998.