Extrakce strukturních informací z běžného textu na základě syntaktického analyzátoru

Miloš Jakubíček

Centrum zpracování přirozeného jazyka, Fakulta informatiky,
Masarykova univerzita, Brno

Úvod

Tento článek popisuje využití syntaktického analyzátoru (parseru) *synt*, který je dlouhodobě vyvíjený v Centru zpracování přirozeného jazyka FI MU Brno, k získání syntaktických (sub)struktur ve vstupní větě.

Jako hlavní výsledek syntaktické analýzy jsou většinou prezentovány derivační stromy popisující odvození dané věty v gramatice používané analyzátorem. Úplná syntaktická analýza češtiny je ovšem stále otevřený problém spočívající zejména v masivní víceznačnosti výstupu: v některých případech může parser pro vstupní větu poskytnout až řádově miliardy výstupních stromů.

Pro některé aplikace ovšem nejsou tyto odvozovací stromy nezbytně nutné, dokonce ani žádoucí – ať už se jedná o extrakci a zpracování informací, převod vět do logické struktury *predikát(argumenty)* nebo o hrubou extrakci slovesných valencí. Ve všech těchto případech očekáváme jiný druh výstupu analýzy poskytující náhled na syntaktické struktury vyskytující se ve vstupní větě. Těmito strukturami jsou zejména jmenné, předložkové a slovesné fráze, jednoduché věty nebo jiné větné segmenty.

Aby však byla taková extrakce oproti běžnému výstupu přínosná, potřebujeme tyto struktury identifikovat *jednoznačně*. Proto jsme upravili syntaktický analyzátor *synt* tak, aby umožňoval extrakci struktur, které z pohledu složkové analýzy odpovídají neterminálům vyskytujícím se v gramatice použité při analýze. Pro zpětné zpřesnění morfologické analýzy (a to až o 30 %) umožňující hlubší rozlišování jednotlivých (sub)struktur byly zapojeny některé dosud nevyužité výsledky analýzy, které jsou popsány dále společně s ukázkami výsledků a příkladem použití pro hrubou extrakci slovesných valencí.

Syntaktický analyzátor *synt*

Syntaktický analyzátor synt (Kadlec – Horák, 2005) provádí analýzu typu chart na základě bezkontextové složkové gramatiky pro češtinu. Aby bylo možné gramatiku jednoduše editovat, jsou její úpravy prováděny na úrovni tzv. metagramatiky (mající přibližně 200 pravidel), ze které je následně vygenerována plná forma gramatiky (obsahující více než 4 000 pravidel). Mimo gramatiku jsou dále udržována pravidla pro tzv. kontextové akce, které pokrývají např. gramatickou shodu.

Nedávná měření (Kadlec, 2007) ukázala, že *synt* dosahuje vysokého pokrytí převyšujícího 94 %, jeho výstup je ovšem zásadně víceznačný, takže počet výstupních stromů se v některých případech pohybuje i v řádu miliard.

Proto analyzátor zahrnuje dvě základní metody, jak se s takto víceznačným výstupem vyrovnat: první z nich představuje rozdělení pravidel gramatiky do různých úrovní a následnou filtraci výstupu podle těchto úrovní; druhá metoda spočívá ve výpočtu pravděpodobností jednotlivých pravidel, ze kterých je pak pro každý strom vypočteno ohodnocení určující pořadí daného stromu ve výstupu analýzy.

Pro účely extrakce struktur byla použita interní datová struktura *syntu*, tzv. chart, multigraf, který je v průběhu analýzy vytvořen a zahrnuje v sobě všechny výstupní stromy. Zásadní na této datové struktuře je její polynomiální velikost (Horák, 2001) umožňující její efektivní zpracování (zatímco počet stromů může být až exponenciální vzhledem k délce vstupní věty a zpracování jednotlivých stromů je tedy výpočetně neproveditelné). Zpracováním chartu zde rozumíme zpracování jeho obsahu po skončení analýzy.

Extrakce struktur

Protože jsou různé syntaktické struktury vzájemně značně odlišné, neexistuje univerzální postup, který by mohl být uplatněn na všechny struktury, a proto bylo nutno vyvinout několik způsobů, jak tyto struktury identifikovat. Zároveň je extrakce prováděna tak, aby byla jednoznačná, tj. snaží se zahrnout všechna možná odvození, která jsou ve výsledném chartu.

Jelikož některé struktury mají rekurzivní podobu (např. jmenné a předložkové fráze), nabízí se jako vhodný výchozí postup extrahovat v takovém případě největší, resp. nejmenší struktury (z pohledu rekurzivity), pro kvalitní výsledky je nicméně nutné tyto postupy kombinovat a rozšířit (např. o dělení struktur podle gramatické shody).

Výsledky extrakce pro různé struktury jsou zobrazeny na následujících příkladech:

• jednoduchá věta (se zanořením):
Vstup: *Muž*, *který stojí u cesty*, *vede kolo*.
Výstup:
```
[0-9): Muž , , vede kolo
[2-6): který stojí u cesty
```

• slovesná fráze:
Vstup: *Kdybych to byl býval věděl*, *byl bych sem nechodil*.
Výstup:
```
[0-5) : byl býval věděl
[6-10): byl bych nechodil
```

• jednoduchá věta (sekvence):
Vstup: *Vidím ženu*, *která drží růži*, *která je červená*.
Výstup:
```
[0-3): Vidím ženu ,
[3-7): která drží růži ,
[7-10): která je červená
```

• jmenná fráze:
Vstup: *Tyto normy se však odlišují nejen v rámci různých národů a států*, *ale i v rámci sociálních skupin*, *a tak považuji dřívější pojetí za dosti široké a nedostačující*.

Výstup:

```
0-2): Tyto normy
6-12): v rámci různých národu a států
15-19): v rámci sociálních skupin
[23-30): dřívější pojetí za dosti široké a nedostačující
```

• jmenná fráze (s dělením podle gramatické shody):

Vstup: viz předchozí věta

Výstup:

```
[0-4): Tyto normy se však
[6-8): v rámci
[8-12): různých národů a států
[13-17): ale i v rámci
[17-19): sociálních skupin
[23-25): dřívější pojetí
[25-30): za dosti široké a nedostačující
```

Literatura:

HORÁK, Aleš: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. Disertační práce. Brno: Fakulta informatiky, Masarykova Univerzita 2001.

KADLEC, Vladimír: Syntactic analysis of natural languages based on context-free grammar backbone. [Disertační práce.] Brno: Fakulta informatiky, Masarykova Univerzita 2007.

KADLEC, Vladimír – HORÁK, Aleš: New Meta-grammar Constructs in Czech Language Parser *synt*. In: Lecture Notes in Computer Science. Berlin – Heidelberg: Springer 2005, s. 85 – 92.