Oponentský posudek

Práce pana bc. Josefa Čecha se zabývá automatickým rozpoznáváním segmentů a klauzí v českých souvětích.

Předkládá rozbor problematiky a několik postupů pro rozpoznávání segmentů a následně klauzí, opatřených četnými pokusy a podrobným kvantitativním vyhodnocením. Staví především na výzkumu doc. Kuboně a doc. Lopatkové, kteří pojem segmentu a možnosti jeho využití pro pomoc automatické syntaktické analýze poprvé představili.

Výzkum je motivován předpokladem, že segmentaci je možné učinit s vysokou přesností jen na základě morfologického označkování, rozdělit tak souvětí na kratší, samostatně zpracovatelné celky, a vylepšit tím celý proces automatické syntaktické analýzy co do úspěšnosti a výpočetní náročnosti.

Tento předpoklad je dobře fundovaný a jeho dosavadní nevyužití se jeví jako překvapivé a zarážející. Striktní interpunkční pravidla a velmi omezený počet tokenů, které se mohou vyskytovat mezi segmenty, jakož i silná autonomie klauzí v mnoha ohledech naznačují, že neprozkoumat tuto cestu by bylo trestuhodné.

Diplomová práce obsahuje první experimenty s aplikací segmentace na nalézání klauzí. Prezentuje pravidlový algoritmus segmentace, který určí všechna přípustná schémata "zanoření" segmentů. Dále pravidlový a statistický algoritmus pro určení klauzí na základě možných schemat a také jejich kombinaci, která se jeví nejúspěšnější. V neposlední řadě navrhuje autor sadu pozičních tagů pro segmenty, kterou sám využívá.

Za touto diplomovou prací je práce opravdu požehnaně. Zadání "spojování segmentů" bylo implementováno třemi různými způsoby, softwarová distribuce hladce na mém počítači zreprodukovala experimenty, označkovala moje data a umožnila mi výsledky vizualizovat.

Práce má však taktéž mnoho nedostatků.

- Největší slabinou celého díla je, že se nepodařilo ověřit výše zmíněný předpoklad. Celá práce je motivována pomocí automatické syntaktické analýze. Ovšem všechny vyvinuté algoritmy pro identifikaci klauzí uhodnou jedno snad správné rozdělení a v tomto uhodnutí je zdrcující chybovost. Jestliže v absolutní většině vět, kde je klauze se vsuvkou, je chyba, jak může takové předzpracování být základem k další syntaktické analýze?

  Nepochybuji o tom, že se autor pokusil dosáhnout úspěšnosti co nejvyšší. Avšak tváří v tvář těmto číslům měl vytvořit nástroj, který by dal na výstup všechna přípustná rozdělení na klauze. Tak by se nechaly otevřené dveře dalšímu zpracování.

  K diplomantově cti je nutno podotknout, že chybové konstrukce sám vytypoval a popsal a ohodnotil. To je znak dobrého výzkumu a byla to práce nad rámec povinného pemza.

- Nízkou úspěšnost vykazuje především statistická metoda. Autor popisuje mimo jiné metodu, kde nevalidní výstup statistického algoritmu je pravidlově opravován do validního stavu a tím se dosahuje vyšší úspěšnosti. Povaha porušení validity naznačuje, že omezující podmínky byly velmi chabě zahrnuty do modelu. Kdyby byl statistický model vhodněji zvolen, byla by chybovost možná podstatně menší. Na ÚFALu je dostatek odborníků na statistické modelování. Autor měl vyhledat jejich konzultaci.

  Na druhou stranu využití segmentů pro detekci klauzí je úloha, kde větší potenciál vidím v pravidlovém přístupu, proto nedostatky ve statistickém zpracování nejsou v tomto případě příliš závažným prohřeškem.

Kromě těchto vážných nedostatků je k nalezení množství dalších, méně podstatných.

- Jako problémová konstrukce se uvádí vsuvka ohraničená pomlčkami. Podle příznaků v povrchové struktuře jazyka lze stěží určit, zda pomlčky opravdu ohraničují vsuvku. Proč tedy autor tyto konstrukce zkrátka neignoruje? Jestliže se má výstup segmentace použít jako vstup pro parser, je precision řádově důležitější než recall. Cílem je rozdělit větu s co největší jistotou a zbytek nechat na parseru. Jestliže tady tedy jistotu nemám, proč nedat ruce pryč?

- V sekci 5.2 se píše: "Analýzu částí uvnitř párových oddělovačů je možné oddělit od zbytku věty a provést ji zcela samostatně." Tvrzení není ničím podloženo.
- V úvodu stojí, že parsery mají nižší úspěšnost na delších větách. To nechci napadat, ale v práci je to podložené jenom jedním článkem, který je deset let starý a vyšel ještě před MST a dalšími inovacemi. Vzhledem k tomu, že na tomto postulátu je založena celá motivace pro tuto práci, bylo by dobré podložit to něčím novějším (nebo udělat vlastní analýzu).
- Ve slovníčku pojmů se definuje segment jako "lingvisticky motivovaná jednotka", pak se uvádí intuitivní vysvětlení a potom citace, že segment je od separátoru k separátoru. A separátor se potom definuje jako to, co odděluje segmenty.

  Považoval bych za šťastnější definovat separátory pro češtinu (ty mají jasnou definici výčtem / tagem) a potom na základě nich definovat segment. Zredukovala by se tím vágnost i délka textu.

  Podobně klauze je definovaná jako "neprázdná množina segmentů na jedné úrovni zanoření mezi dvěma hranicemi klauzí." Přitom hranice klauzí je definována jako "hranice segmentu, která zároveň uvozuje novou klauzi nebo ukončuje předchozí."

  Autor správně uvádí, že čistě na základě povrchové struktury jazyka nelze klauzi přesně definovat. Ani to ale podle mě není potřeba a tyto bludné definice vůbec nemusely vzniknout. Klauze jsou součástí jazyka. Jejich definice je úkolem lingvistů. My je pozorujeme a chceme je nalézt. Zanoření segmentů je pak dáno vztahem přítomných klauzí.
- V tabulkách 14a a 14b stojí podíl správně určených celých vět co do hloubky zanoření segmentů 71% resp. 56%. Tabulka 15a pak uvádí podíl správně určených celých vět co do detekce klauzí 87%. Není mi jasné, jak může být více vět, kde všechny klauze jsou správně určené, než těch, kde je správně určeno schéma zanoření segmentů.
- V práci stojí, že ve výstupu segmentačního algoritmu určitě bude správné schéma zanoření. Chybí k tomu podložení. Protipříklady, které jsem vymyslel, jsou zajisté obskurní, ale přesto poukazují na autorovu neopatrnost.

  Protipříklady proti bodům na str.17 (sekce 5.4)

  - "Že jsem šel sám, jsem přesvědčen, že moje matka věděla." první segment má úroveň 2 (bod 1)
  - "Věděl, když půjde po tmě, že ho neuvidí." druhý segment je o dvě úrovně níž než první (bod 3)
  - "Chceš vědět, Ludvíku, můj starý - a dalo by se říci - ano! Vlastně jediný dobrý příteli, jak jsem se zde v této osamělé pustině uprostřed hlubokého lesa, kde široko daleko nikdo nebydlí a kudy nevede ani pěšina a kde bys za celý den nespatřil človíčka, nebýt těch zatracených vzducholodí, ocitl?" (z Hospody na mýtince) vykřičník neukončuje větu (bod 5)
- Na straně 16 se hovoří o "této situaci" a "podobných konstrukcích" a odkazuje se na obrázek, ale není zcela zřejmé, o čem je vlastně řeč.
- Na straně 32 se zmiňuje gerundium. Výskyt tohoto jevu v češtině mi není znám. Snad autor myslí přechodníky?
- Jsou přítomny jazykové chyby. Anglický abstrakt je napsán kostrbatou angličtinou.

Závěr

Jakkoliv četná a významná jsou místa pro zlepšení a další rozvoj, diplomová práce pana bc. Čecha svým rozsahem a provedením více než bohatě splňuje nároky na diplomovou práci v jejím zadání. K obhajobě ji doporučuji.