Posudek doktorské disertační práce ing. Pavla Kopřivy nazvané

Automatická identifikace strukturních korespondencí v paralelním korpusu

Předložená práce se zabývá tou oblastí matematické lingvistiky, jejímž předmětem je zpracování paralelních textů přirozeného jazyka. Jejím hlavním výsledkem je návrh algoritmu, jak nacházet odpovídající si strukturní části ve dvojici vět, z nichž jedna je překladem druhé. V práci však nejde o hledání korespondencí na úrovni strukturně nejnižší, tj, na úrovni slovních tvarů ve větách obsažených (i to je ovšem nesmírně obtížný úkol sui generis), nýbrž na úrovních vyšších – na úrovni netriviálních syntaktických struktur, tj. frází či jejich částí, jako jsou například jmenné fráze, slovesné fráze apod. Hledání strukturních korespondencí tedy zahrnuje tři dílčí úkoly: a) identifikovat fráze ve vstupním jazyce, b) identifikovat fráze ve výstupním jazyce a pak c) stanovit, že dané frázi F_Inp ve vstupní větě odpovídá ve výstupní větě fráze F_Out. Autor si za konkrétní jazyky zvolil češtinu a angličtinu a na nich také představil svou metodu.

Problém se v zásadě řeší třemi způsoby: statisticky na základě větších množství trénovacích dat, nebo pomocí syntaktických pravidel platných pro vstupní i výstupní jazyk, případně i kombinací uvedených způsobů. Autor zvolil pracnou metodu založenou na pravidlech v naději, že se mu podaří přesněji vystihnout syntaktické struktury v obou zvolených jazycích i jejich korespondenci, než by to dokázala metoda statistická, svou povahou vždy nepřesná.

Autor operuje s pojmem chunk jakožto pojmem centrálním. V průběhu svého doktorského studia vyvinul desítky pravidel identifikujících v obou jazycích hlavní chunky i jejich korespondence, vycházeje přitom z předpokladu, že už má nějakým způsobem zarovnána slova; příslušný algoritmus slovního zarovnání sám nevyvinul, použil tu dostupných nástrojů. Je jasné, že jsou-li jednotlivé slovní tvary zarovnány chybně, projeví se to negativně i na vytváření korespondujících si chunků, jejichž jsou jednotlivá slova jakožto jakési elementární chunky součástí.

V autorově pojetí je důležité, že jeho chunky mohou mít hierarchickou strukturu, tj. chunk se může skládat z menších chunků. V ideálním případě lze dosáhnout zpracování celé věty, jejíž struktura je největším možným chunkem zahrnujícím všechna slova věty.

Ústřední, tvůrčí část práce tvoří kapitoly čtvrtá až šestá, kde autor popisuje svou metodu založenou na pravidlech včetně její počítačové implementace. Vyvinul svůj vlastní jazyk pro zachycení pravidel, syntaktická pravidla také sám napsal a navíc je rovněž implementoval a vyzkoušel na několika dokumentech, zejména knihách: vstupem byl vždy text anglický, překladem text český. V posledních dvou kapitolách uvádí dosažené výsledky a jejich přehled, i

zvolené míry. Výsledky vhodně interpretuje včetně zdůvodnění, proč ty či ony struktury byly zpracovány více nebo naopak méně úspěšně. Vedle jasných, jednoduchých případů uvádí podle vzrůstající složitosti jednotlivé typy problémů exemplifikované příslušnými větami a náležitě zdůvodňuje větší či menší úspěšnost jejich zpracování.

Práce ukazuje, nakolik je nacházení korespondujících si struktur obtížné, protože zvolené jazyky, ač oba patří k indoevropské rodině, jsou typologicky velmi odlišné. Ukazuje se, že statistický přístup, následně nijak ručně nekorigovaný, vede k poměrně velkému množství nesprávných korespondencí, proto také autor bohužel nemohl vycházet z nějakých správných vzorových dat, s nimiž by úspěšnost své metody mohl srovnávat. Metoda založená na pravidlech, kterou autor zvolil, je, zdá se, nadějnější, to by se však jasněji vyjevilo poté, co by pravidel bylo více než jen ke dvěma stovkám (samozřejmě za předpokladu správné elementární korespondence na úrovni slov). Je jasné, že čím více – pochopitelně správných – pravidel, tím by byly výsledky lepší. Odhaduji, že síla pravidel by se projevila, až by jich bylo alespoň kolem pěti set (jak autor uvádí pro srovnání, pravidel pro automatickou morfologickou disambiguaci bylo dosud vyvinuto dvojicí pracovníků kolem 2500 v období cca 13 let) a tento základ obsahující zejména obecná pravidla (čili nikoli lexikálně závislá) by se pak mohl postupně rozšiřovat právě o lexikálně specifická pravidla, ustálená slovní spojení apod.

Autor chtěl ověřit kvalitu a smysluplnost své metody vývojem 162 pravidel, jejich implementací a aplikací na reálných datech. To se mu, myslím, podařilo. Ukazuje se, že tato pravidla fungují za předpokladu správných vstupních podmínek (zejména zarovnání na slova) pozoruhodně dobře, zvláště – pochopitelně – na strukturách v obou jazycích podobných. Lze se tak kojit velkou nadějí (jak autor doložil na vytvořených pravidlech), že právě rozšířením dosavadního repertoáru pravidel lze dosáhnout lepších výsledků. Situaci ovšem objektivně komplikuje často podceňovaná, ač zřejmá skutečnost, že angličtina a čeština se liší více, než se na první pohled soudí, a pravidel tedy musí být hodně.

Závěrem: mám za to, že disertantova práce je dobrým metodologickým základem pro zpracování syntaktických stukturních korespondencí v odpovídajících si větách paralelního korpusu. Říkám metodologickým, protože zpracování konkrétní dvojice jazyků pochopitelně vyžaduje osobitá pravidla specifická pro příslušné jazyky. Práci doporučuji k obhajobě: doktorand totiž podle mého soudu prokázal, že dokáže samostatně přemýšlet a vědecky pracovat.

V Praze dne 1. 4. 2013

doc. RNDr. Vladimír Petkevič, CSc., školitel