Posudek disertační práce (Alexandr Rosen)

RNDr. Ondřej Bojar
Exploiting Linguistic Data in Machine Translation

Deklarovaným tématem práce jsou vztahy mezi lingvistickými teoriemi, [jazykovými/lingvistickými] daty a aplikacemi. Konkrétně autor zkoumá, jaký vliv má využití teoretické koncepce FGP (funkční generativní popis) a lexikálních dat, zejména valenčních rámců sloves, na kvalitu strojového překladu. Lexikální data přitom mohou být vytvořená lexikografy nebo získaná z korpusu. První třetina práce (kapitola 2) zkoumá právě metody extrakce valenčních rámců z textů. Další část (kapitola 2) se věnuje experimentům se systémy statistického strojového překladu, které provádějí transfer na lingvisticky motivovaných rovinách s různou mírou abstrakce, včetně nejvyšší, tektogramatické roviny. V poslední třetině pak autor popisuje metodu „frázového" překladu, která však s lingvistickými rovinami popisu explicitně nepracuje. Ani v jednom případě není přínos lingvistické teorie a lexikálních dat jednoznačně pozitivní.

To ani zdaleka neplatí o přínosu této práce. Od začátku až do konce autor udržuje lať ku nasazenou hodně vysoko, a to ve všech možných disciplínách, počínaje zcela přirozeným a srozumitelným jazykem (působícím jako projev rodilého mluvčího) a téměř stoprocentní formální dokonalostí, přes velice bohaté odkazy, citace a argumenty, až po metodologicky bezchybnou kostru, vyvozující z poctivě získaných dat nezpochybnitelné, jakkoli nejednoznačné závěry. Tak například otázku užitečnosti slovníku ve vztahu k systémům zpracování přirozeného jazyka autor řeší nejen ve stěžejních kapitolách, ale i v závěrečné diskusi (kapitola 5), a to výstižným shrnutím závěrů různých autorů. Současnou úroveň strojového překladu ilustrují příklady v příloze, z nichž i ten nejúspěšnější je jen obtížně srozumitelný. Přesto je zajímavé porovnávat výsledky jednotlivých systémů i jinak než podle BLEU: např. některé překlady si lépe poradí třeba se shodou.

Mezi velké přednosti práce patří i hledání příčin neuspokojivých výsledků, které autora skoro vždy přivedou k námětům na další práci. Vztah mezi lingvistickými teoriemi a daty na straně jedné a aplikacemi na straně druhé je obvykle složitý, zprostředkovaný, ovlivněný mnoha parametry, jejichž optimální nastavení nelze předem odhadnout. Proto autor tento vztah zkoumá i z druhého konce: přímou (frázovou) metodu úspěšně modifikuje lingvistickými moduly.

Ve vypočítávání předností práce by se dalo dlouho pokračovat. Místo toho odkážu na její závěr (str. 108–110), zejména na výčet dosažených výsledků. Následující seznam představuje jen drobné vady na kráse:

- str. 13, 2. věta:
  [...] *linguistic theories are used as a basis when prescribing what is an appropriate and correct usage of an expression* [...]
  – možná, že se nakonec takto s lingvistickými teoriemi skutečně zachází, ale jejich autoři to tak většinou nemyslí, jejich cílem je jazyk popisovat a hledat v něm nějaký systém, nikoli předepisovat, jak je to správně

- Část 2.2.1, str. 17 a dále: značkování v PZK se někdy liší od teorie FGP, hlavně na analytické rovině. Mělo by se na to upozornit.

- Str. 19, 3. odstavec: *the t-layer includes nodes for entities that were not explicitly expressed* [...] *This is one of several reasons that make the t-layer language dependent and not an Interlingua*. Není mi jasné, proč by právě doplnění nevyjádřených větných členů znemožňovalo považovat tektogramatickou rovinu za interlingvu. Prosím o vyjasnění.

- Str. 21, tabulka 2.1: SYN2000 *je* vyvážený korpus, aspoň podle http://ucnk.ff.cuni.cz/syn2000_bonito.html: *Tento korpus je vytvořený z celých textů, které do něj byly zařazeny na základě výzkumů recepce psaného jazyka tak, aby pokrývaly co nejširší žánrové rozvrstvení češtiny.*

- Str. 23, 2. odstavec od konce: Pro nepoučeného cizince by bylo dobré vysvětlit, co je to ACT(1), PAT(4), DIR3(*). Také mu nemusí být jasné, kde a jak hledat větu `t-cmpr9410...`

- Str. 43, část 2.9.2 Verbs of Communication: některá ze sloves umožňují vyjádřit „informaci" kromě vedlejší věty také infinitivem. Proč by to nešlo využít?

- Str. 54, 1. věta: *In essence, the task of MT is to efficiently store and correctly reuse pieces of texts previously translated by humans to translate sentences never seen so far.* I když se dále píše o tom, že metody založené na pravidlech *operate with a very distilled representation of words and their translations*, výše uvedená charakteristika úkolu strojového překladu obecně neobstojí. Tradiční systémy založené na pravidlech s dříve přeloženými texty často vůbec nepracují a celý proces překladu řeší na základě implementovaných gramatik příslušných jazyků a modulů transferu, které nebyly extrahovány z textů, ale vytvořeny ručně.

- Str. 57, část 3.1.3: seznam výjimek z volného slovosledu v češtině by se dal rozšířit třeba ještě o relativně pevný slovosled ve jmenných skupinách.

- Str. 97 a dále, část 4.8: chvíli trvalo, než oponent pochopil, co jsou to *verb-modifier relations*. Při první zmínce (odstavec 2) by to možná chtělo příklad nebo synonymum.

- Stejný odstavec: *local agreement (within 3-word span) is relatively correct but...* Jediný příklad, kdy okénko tří slov nevyřeší morfologický problém, je ten výše uvedený. Jak je to třeba se shodou podmět-přísudek? Obr. 7.2 naznačuje, že výsledek není zaručen, i když se podmět i přísudek vejdou do okénka: *firmy vyběhl*. A co když větné členy, u kterých se shoda/valence morfologicky projevuje, stojí dál od sebe? Je to o hodně horší?

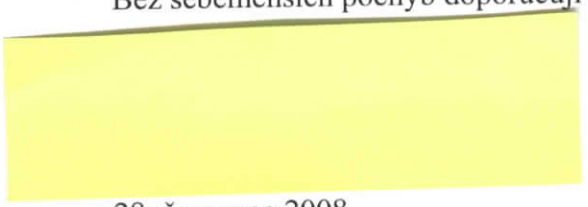
Překlepy apod.:

- str. 36, část 2.7.2, řádek 1: má být *WFD* místo *WSD*
- str. 38, „Algorithm 1", bod 3: dtto
- str. 39, „SIMPLE", poslední věta: chybí tečka před závorkou
- str. 50, část 2.10.5, 1. odstavec: *the the aim*

- str. 96, poslední věta předposledního odstavce: má být *which suggests* místo *which suggest*
- str. 98, 5. řádek od konce: chybí čárka za *(lexicalized)*
- str. 105, poznámka 5, 2. věta: *the WSD module is not used a feature* – ?

Závěr

Úkol, který si autor zvolil – prozkoumat vztahy mezi lingvistickou teorií, daty a aplikacemi – pochopitelně přesahuje možnosti jedné disertace. Omezení na jednu konkrétní teorii, lexikální data s valenčními rámci a strojový překlad téma významně zužuje, ale jinak by práce nutně zůstala příliš obecná. Ve zvolené podobě představuje významný přírůstek do seznamu doporučené četby pro zájemce o strojový překlad, ale i o obecnější otázky využití lingvistické teorie a lingvistických dat v počítačových aplikacích. (Čímž budiž řečeno, že publikace by byla žádoucí.) Práce jednoznačně prokazuje předpoklady autora k samostatné tvořivé práci. Bez sebemenších pochyb doporučuji udělení titulu Ph.D.



28. července 2008