Vyjádření vedoucího doktorské disertační práce

Student: Mgr. Michal Novák

Název práce: Coreference from the Cross-lingual Perspective

Školitel: doc. Ing. Zdeněk Žabokrtský, Ph.D.
Institute of Formal and Applied Linguistics
MFF UK
Malostranské náměstí 25
118 00 Praha 1

Popis:

Hlavním cílem předložené práce je zkoumání možností automatické analýzy koreferenčních vztahů v situaci, kdy lze pracovat s daty více jazyků najednou.

Práce má standardní strukturu. Po úvodní kapitole představující cíle práce následuje kapitola, která shrnuje základní pojmy a podrobněji předkládá klasifikaci koreferujících výrazů. Třetí kapitola shrnuje existující základní monolingvální metody rozpoznávání koreference, a dále již publikované metody pro rozpoznávání koreference s využitím dat v jiných jazycích. Kapitola 4 popisuje existující datové sady relevantní pro pár angličtina-čeština. Kapitola 5 podrobně analyzuje podobnosti a rozdíly v manifestaci koreferenčních vztahů mezi češtinou a angličtinou na základě pozorování autentických paralelních dat. Následují tři původní experimentální kapitoly. Kapitola 6 se zabývá úlohou párování koreferenčních výrazů napříč paralelními daty. Kapitola 7 využívá pro zkvalitnění monolingváního rozpoznávání koreferenčních vztahů rysy přenesené z druhého jazyka (na základě zarovnání v paralelním korpusu). Kapitola 8 se zabývá scénářem, kde anotace koreference pro jazyk s nedostatkem datových zdrojů vznikne projekcí z jiného jazyka, pro který taková data existují. Závěrečná devátá kapitola shrnuje dosažené výsledky.

Hodnocení:

Zhruba v poslední dekádě lze zaznamenat zvýšený zájem o „multilingvální“ přístup k některým úlohám z počítačové lingvistiky. Primární motivací obvykle bývá „*data acquisition bottleneck*“ (tj. problém s nedostatkem konkrétních lingvisticky relevantních dat pro jednotlivé jazyky) a snaha o jeho alespoň částečného zredukování díky využití nějaké formy přenosu anotované informace napříč jazyky, za jazyků „bohatších“ na zdroje do jazyků „chudších“.

Úloze koreference zatím bylo v tomto multilingválním nastavení věnování relativně málo pozornosti, na rozdíl například od syntaktické analýzy, kde standardizace datových zdrojů v jednotlivých jazycích (jakožto základní podmínka multilingválních experimentů) a vývoj metod přenositelných mezi jazyky (včetně soutěží, tzv. shared tasks) mají již více než desetiletou tradici. Z tohoto pohledu je koreference dosud Popelkou a žádnou svoji obdobu Universal Dependencies ani vzdáleně nenabízí. V tomto kontextu tedy považuji předloženou práci za velmi aktuální a originální.

Další rozdíl, který okamžitě vyvstane při srovnání se závislostní syntaktickou analýzou v prostředí vícejazyčných dat, spočívá v tom, že samotná úloha koreference má od začátku složitější strukturu. Například (i) není předem zjevné, která slova se účastní koreference, (ii) koreferenčních vztahů je více typů a některé z nich interagují se syntaktickou strukturou věty, (iii) není jasné, jak

vyhodnocovat úspěšnost rozpoznání koreference, protože koreferenční vztahy spolu tvoří větší celky.

Domnívám se, že Michal prokázal, že do zkoumané problematiky získal hluboký vhled. V předložené práci pak spatřuji dva hlavní přínosy. Zaprvé, předvedl precizní analýzu typových rozdílů ve vyjadřování koreference v češtině a angličtině. Zadruhé, a to je podle mého názoru těžiště práce, navrhl, realizoval a pečlivě vyhodnotil dva úspěšné experimenty, které ukazují, jakými metodami lze anotovanou informaci o koreferenčních vztazích využít napříč jazyky.

Některé z výsledků byly publikovány již v minulosti, a převážně formou konferenčních článků, celkově jich bylo publikováno cca 20.

Domnívám se, že po jazykové i formální stránce je text zpracován kvalitně a splňuje obvyklá kritéria na doktorskou disertační práci.

Závěr:

Předloženou práci považuji za originální a pečlivě zpracovaný vědecký příspěvek k tématu analýzy koreference a doporučuji ji k obhajobě.

Jahodov, 31. srpna 2018

doc. Ing. Zdeněk Žabokrtský, Ph.D.