СТАНОВИЩЕ

от доц. д-р Палмира Легурска

за дисертационен труд за придобиване на образователната научната степен „доктор"

Автор на дисертационния труд: Георги Емилов Илиев

Тема на дисертационния труд: „Езиково мотивирана оптимизация на машинния превод"

Научен ръководител: проф. д-р Светла Коева

Георги Емилов Илиев е бакалавър по английска филология, СУ „Кл. Охридски" – 2002 г., бакалавър по скандинавистика (семестриално завършил през 2008 г.), магистър по компютърна лингвистика и интернет технологии в хуманитаристиката – 2009 г., докторант и специализант към Секцията за компютърна лингвистика в ИБЕ БАН.

С решение на Научния съвет на ИБЕ съм определена да участвам в научното жури за осигуряването на процедура по защита на представения от Г. Илиев дисертационен труд за придобиване на образователната и научната степен „доктор" по специалността „Общо и съпоставително езикознание (компютърна лингвистика).

Докторантът е представил материали на електронен и хартиен носител за процедурата – дисертационен труд, автореферат на дисертационния труд, автобиографична справка, списък и текстове на публикациите по труда.

Изброените документи дават представа за Г. Илиев като квалифициран специалист по английска филология, скандинавистика, българска филология – компютърни и интернет технологии. Докторантът е приложил също така сертификати за завършен курс по „Увод в статистиката" и от школа по логика, езици и информатика, допълващи изброените по-горе квалификации. Всичко изброено до тук характеризира докторанта като специалист в областта на компютърната лингвистика и езиковите технологии.

Дисертационният труд е с общ обем 208 с., в това число увод (7 – 54 с.), две глави (55 – 189 с.), заключение (189 – 192 с.), приложение – експериментални корпуси и резултати на електронен носител (193 – 196 с.) и библиография (197 – 208 с.).

Авторефератът е с обем 36 с.

Предвид на това, че в дисертацията се разработват проблеми в междинната област на математическата, компютърната лингвистика и съпоставителното езикознание, ще се спра по-подробно на въпроси от последния модул, който е моя конкретна специалност.

Уводът на работата /с. 7 – 13/ е посветен на обекта и предмета на изследването, целите и задачите, прави се обзор на състоянието на научните изследвания в дадената област, определящи методите на анализ.

Предмет на работата е машинният превод на български език, осъществен чрез фразов статистически метод. Изтъква се, че технологиите за машинен превод се доминират от големите езици – английски, китайски, арабски – и изборът на езиковата двойка – шведски - български език оформя иновативния характер на анализа.

Предметът и целите на изследване са поставени в съответния исторически контекст на разработване на проблематиката и фокусът на внимание е в теоретично-приложен план: посочва се липсата на големи масиви от паралелни тестове като обект на анализ и така езици с богата степен на морфологично разнообразие се изучават предимно в условията на ограничени паралелни езикови корпуси, което е безспорен недостатък.

Първият етап от изследването е построяване на работеща система за статистически машинен превод между езици с относително богата степен на морфологично разнообразие в условията на ограничени паралелни езикови ресурси.

Вторият етап се прави с оглед на конкретните типологически характеристики на изследваната езикова двойка: да се потърсят езиково зависими начини за оптимизиране

на построената система за превод от шведски на български език. Системата се обогатява с лингвистична информация от свободно достъпни едноезикови ресурси.

Първа глава /с. 13 - 54/ представя машинния превод в исторически план на изучаване като зона на засилен обществен интерес от 1949 г. и до днес, в която се изучават отделните постижения в американската, руската, европейската и българската школа. Подчертава се необходимостта да се достигне до скрит "универсален език", който да улесни превода между естествените езици. Това е език посредник, притежаващ всички свойства на изследваните езици, със свой речник, морфология, синтаксис. Авторът се базира на модела на комуникационния канал с шум и скритите марковски модели.

Търсят се програмни инструменти за лингвистични изследвания, които да позволят на лингвистите да пишат със собствен код, а не да разчитат на случайно обучени програмисти за съставяне на програми на асемблиращ (сглобяващ) език. Работата с механични средства се сблъсква с онова, което докторантът нарича семантична бариера.

Специално се изброяват шест подхода към построяването на език посредник: *генетичен* /група сродни езици/; *корелационен* /сбор от множествата на всички входящи езици/; *опростяващ* /опростяване на някой от входящите езици до степен, в която той става пригоден да влезе в качеството на език посредник/; *логически* /входящ текст, представен със средствата на математическата логика/; *минималистичен* /множество граматични елементи, образувано като сечение от всички множества от граматични елементи от езиците в полето/; *синтетичен* /езикът посредник се съставя с оглед на общите за дадена двойка езици елементи/. На тази база се обсъждат постиженията на някои емпирични и теоретични направления в Европа — Кеймбриджка група, лингвистична група от Карловия университет в Прага, критиките на Бар-Хилел, Комисията по автоматична обработка на езика и пр. Предмет на специално внимание са разработките, посветени на машинния превод в България и се очертават характерните постижения.

Съвременните системи за машинен превод – езиково зависими преводи (директни, трансферни и интерлингвални), корпусните методи (паралелни корпуси - машинен превод по аналогии , IBM Candide, хибридни преводи) са отделен аспект на изучаване и коментар.

Оформя се общ извод, че по-голямата част от разработките, свързани с машинния превод имат сходни проблеми,определени от естествените ограничения, наложени от обработката на естествените езици с формални методи.

Като споделя мнението на К. Пиотровска, докторантът формулира хипотезата (с.41), че само лексиконът (думи и фрази) в езика, без да се навлиза в синтактичната структура на изречението, може да даде достатъчна информация за разбирането на общия контекст. Това важи в по-голяма степен за синтетичните езици. В човешката реч, както и в повечето печатни текстове се наблюдават синтактични, стилистични и семантични грешки и целостта им се нарушава от повторения и колебания. Поради тази причина анализът на информационните и статистическите свойства на езика показва, че съставянето на текстове може да се представи като сложен марковски процес, който предполага, че текстът се поражда единица по единица, подчинен на концепциите на вероятностната функционална граматика, при която регулярната структура не е задължително признак за наличие на регулярни предопределени релации; строгото планиране и точното предугаждане на структурата на текста по генеративни начини дава резултати от текста, а по-отдалечените езикови единици демонстрират бързо отслабващи стохастични връзки.

Докторантът изтъква, че използването на наличните езикови ресурси и популяризирането на софтуерни решения с отворен код създава възможност да се

интегрира българският език в глобалната мрежа и да се търсят възможните методи за усъвършенстване на машинния превод.

Глава втора е посветена на създаването на система за статистически машинен превод с български като целеви език. Извършва се съотнасяне на паралелни корпуси по изречения, прилага се процедура за статистически модел за пословен превод и статистически модел за фразов превод. Предлага се базова система за фразов статистически превод на български език от логаритмично-линеен характер върху материал от превод от шведски на български език. Оформя се изводът, че независимо от напредъка на фразовите модели за машинен превод, граматичността в близък контекст остава синтагматична. В този смисъл резултатите от статистическия машинен превод зависят от размера на пространството за търсене в езика-цел. Улеснено е декодирането при превод от езици със силно изразени морфологични характеристики на такива с по-малко морфологично вариране. Преводът между езици със сравнително богата морфология се затруднява. При превод от шведски на български език се увеличава броят на непознатите думи, тъй като основните теоретични допускания, използвани за статистически машинен превод, се базират върху преводи на английски език, който е език със сравнително малко лексикални варианти за разлика от езиците, с които докторантът провежда експеримент. Това мотивира съдържанието в трета глава на дисертацията.

Глава 3 „Езиково зависими компоненти и методи" описва построяването на система за статистически машинен превод за различни двойки езици. Като се установява съществена разлика при оценката на качеството на превода, направено с автоматична метрика при превод от английски и шведски на български език, се изказва хипотезата (с. 115), че методите за построяване на базова система за машинен превод дават преимущество за превод между езици с определена типологична характеристика, каквато българският език като целеви не притежава. Проведените експерименти показват най-висока оценка при преводите на шведско-английската система. Затова целта на описваната в тази глава експериментална работа е търсенето на практически възможни езиково зависими методи и компоненти, с които да бъдат обогатени отделните етапи от обучението и работата на системата за статистически машинен превод с цел да се повиши оценката на резултата при превод на български език.. В главата се съдържат осем езикови експеримента: съотнасяне на паралелни корпуси от изречения по /слабо/езиково зависим метод, формално разширяване на изходната част от тренировъчните данни по езиково зависим метод, съкращаване на пространството за търсене и съотнасяне на думи, морфологична информация по модел IBM 4, синтактично пренареждане, езиков модел от морфологични етикети, фактории модели и техните ограничения, нормализиране на резултата на езиково зависим инструмент.

В заключението на работата се изтъква, че в рамките на съвременните системи за машинен превод няма императивно средство с общо приложение. Използването на корпусни решения в машинния превод създава възможности за оптимизацията му. Става възможно да се направи базова система за машинен превод на произволна двойка езици, свързана с достатъчно големи паралелни корпуси.

Глава 2 представлява основа за създаване на собствена изследователска система за фразов статистически машинен превод на български по отделни модулни стъпки, от които се състои системата от съотнесени паралелни данни по изречения, установяването на стъпките, които могат да бъдат обогатени с езиково зависима информация.

Извеждането на приносните моменти в дисертацията оценявам като важна съставка на изследването.

Наличният исторически преглед на проблематиката, придружен от критическия анализ от страна на докторанта, очертава изследователския подход на Г. Илиев като задълбочен и премерен. Търси се баланс между отделните школи и техните постижения, като се очертава мястото и на българските изследователи. Тази част от дисертацията може да се разглежда като отделна разработка, която представлява интерес за лингвистичната общност като отделна публикация. Също така популяризирането на вероятностните модели в обработката на естествения език могат да бъдат основа на бъдещи взаимодействия между отделните изследователски колегии в аспект на перспективното им сътрудничество.

Въпрос буди дали изводите за отделни двойки езици могат да бъдат съотносими и дали не се търсят интуитивно намерени адекватни решения за всяка от тях.

Извлечените експериментални двуезични корпуси: шведско-български, немско-български, датско-български, нидерландско-български, служещи за обучение на базовата система, представляват вероятен принос за съпоставителните изследвания и частните теории на превода.

Обсъдените експериментални методи от гл. 3, за сложните думи в шведския език, пренареждането на словореда, нормализацията на съгласуването в групата на съществителното име са ценни от приложна гледна точка. Съществено е да се формулира какво и как се наблюдава, и какво ни дава това за формулираната задача на експеримента.

Във всички случаи смекчаването на парадокса между естествения и компютърния език е въпрос на бъдещи изследвания и дадената дисертация спомага съществено за това.

Предмет на обсъждане още може да бъде дали корпусните данни, които са продукт на определено приложена лингвистична теория, подлежат на концептуална симетрия или е необходимо езиковите данни да бъдат обработвани от гледна точка на единна методика.

Представените публикации (три самостоятелни и две в съавторство), както и трите участия на автора в семинари с презентации, допълват и разгръщат проблемите, обсъждани в дисертацията.

Представеният реферат отразява точно и ясно съдържанието на дисертацията.

Дисертационният труд е посветен на актуален проблем от компютърната лингвистика, използвани са съвременни методи и техники за анализ на данните и в този смисъл работата има оригинален и приносен характер.

Давам положителна оценка на предложения дисертационен труд и смятам, че научното жури може да присъди на Георги Емилов Илиев образователната и научната степен „доктор".

Октомври 2014 г.

Доц. д-р Палмира Легурска