

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ СОЦИОКУЛЬТУРНЫХ КОММУНИКАЦИЙ
Кафедра компьютерной лингвистики и лингводидактики

Реферат дипломной работы

СОЗДАНИЕ И ОБРАБОТКА ПАРАЛЛЕЛЬНОГО КОРПУСА ТЕКСТОВ
ДЕЛОВОЙ ТЕМАТИКИ ДЛЯ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА НА
НЕЙРОННЫХ СЕТЯХ

Даргель Наталья Вячеславовна

Руководитель Воронович Валерий Викторович

Минск, 2021

РЕФЕРАТ

дипломной работы Даргель Н.В.

СОЗДАНИЕ И ОБРАБОТКА ПАРАЛЛЕЛЬНОГО КОРПУСА ТЕКСТОВ ДЕЛОВОЙ ТЕМАТИКИ ДЛЯ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА НА НЕЙРОННЫХ СЕТЯХ

Дипломная работа включает 74 страницы, 23 рисунка, 1 таблицу, 4 приложения и основана на 44 источниках.

Ключевые слова: МАШИННЫЙ ПЕРЕВОД, НЕЙРОННЫЕ СЕТИ, КОРПУС ТЕКСТОВ, ПАРАЛЛЕЛЬНЫЙ КОРПУС, АВТОМАТИЗИРОВАННЫЙ ПЕРЕВОД.

Цель исследования – выявить роль корпусов текстов в системах машинного перевода на нейронных сетях.

Объект исследования – системы машинного перевода.

Предмет исследования – корпус текстов в составе систем машинного перевода на нейронных сетях.

Актуальность исследования обусловлена отсутствием качественной системы русско-белорусского делового перевода, а также представленностью кодексов Республики Беларусь только на одном из государственных языков (русском).

Методы, использованные при написании дипломной работы: анализ, описание, синтез, сравнение, структурно-функциональный метод.

Научная новизна состоит в том, что в данной работе впервые на основании практического исследования были сформулированы особенности работы систем русско-белорусского машинного перевода, использующих параллельный корпус текстов.

Практическая значимость работы состоит в анализе и доработке переведенных с помощью машинного перевода Кодекса Республики Беларусь об образовании и Налогового кодекса Республики Беларусь, выявлении ошибок и оценке качества перевода, в создании параллельного русско-белорусского текста для системы машинного перевода на нейронных сетях.

Результаты исследования были апробированы на 77-ой научной конференции студентов и аспирантов БГУ, Минск, 07.04.2020; 78-ой научной конференции студентов и аспирантов БГУ, Минск, 13.05.2021 (выступление с докладом «Создание и обработка параллельного корпуса текстов деловой тематики для системы машинного перевода на нейронных сетях»).

Автор работы подтверждает, что приведенный в работе материал объективно отражает состояние исследуемого процесса, а все заимствования из источников сопровождаются ссылками на их авторов.

)

РЭФЕРАТ

дыпломнай працы Даргель Н. В.

СТВАРЭННЕ І АПРАЦОЎКА ПАРАЛЕЛЬНАГА КОРПУСА ТЭКСТАЎ ДЗЕЛАВОЙ ТЭМАТЫКІ ДЛЯ СІСТЭМЫ МАШЫННАГА ПЕРАКЛАДУ НА НЕЙРОННЫХ СЕТКАХ

Дыпломная праца ўключае 74 старонкі, 23 малюнкі, 1 табліцу, 4 дадаткі і заснавана на 44 крыніцах.

Ключавыя словы: МАШЫННЫ ПЕРАКЛАД, НЕЙРОННЫЯ СЕТКІ, КОРПУС ТЭКСТАЎ, ПАРАЛЕЛЬНЫ КОРПУС, ЛІНГВІСТЫЧНАЕ ЗАБЕСПЯЧЭННЕ, АЎТАМАТЫЗАВАНЫ ПЕРАКЛАД.

Мэта даследавання – выявіць ролю карпусоў тэкстаў у сістэмах машыннага перакладу на нейронных сетках.

Аб'ект даследавання – сістэмы машыннага перакладу.

Прадмет даследавання – корпус тэкстаў у складзе сістэм машыннага перакладу на нейронных сетках.

Аktуальнасць даследавання абумоўлена адсутнасцю якаснай сістэмы руска-беларускага дзелавога перакладу, а таксама прадстаўленасцю кодэксаў Рэспублікі Беларусь толькі на адной з дзяржаўных моў (рускай).

Метады, выкарыстаныя пры напісанні дыпломнай працы: аналіз, апісанне, сінтэз, параўнанне, структурна-функцыянальны метады.

Навуковая навізна складаецца ў тым, што ў дадзенай працы ўпершыню на падставе практычнага даследавання былі сфармуляваны асаблівасці працы сістэм руска-беларускага машыннага перакладу, якія выкарыстоўваюць паралельны корпус тэкстаў.

Практычная значнасць працы складаецца ў аналізе і дапрацоўцы перакладзеных з дапамогай машыннага перакладу Кодэкса Рэспублікі Беларусь аб адукацыі і Падатковага кодэкса Рэспублікі Беларусь, выяўленні памылак і ацэнцы якасці перакладу, у стварэнні паралельнага руска-беларускага тэксту для сістэмы машыннага перакладу на нейронных сетках.

Вынікі даследавання былі апрабаваныя на 77-ай навуковай канферэнцыі студэнтаў і аспірантаў БДУ, Мінск, 07.04.2020; 78-ай навуковай канферэнцыі студэнтаў і аспірантаў БДУ, Мінск, 13.05.2021 (выступленне з дакладам "Стварэнне і апрацоўка паралельнага корпуса тэкстаў дзелавой тэматыкі для сістэмы машыннага перакладу на нейронных сетках»).

Аўтар працы пацвярджае, што прыведзены ў працы матэрыял аб'ектыўна адлюстроўвае стан даследаванага працэсу, а ўсе запазычаныя з крыніц суправаджаюцца спасылкамі на іх аўтараў.

SUMMARY

of diploma paper by Dargel N.V.

CREATING AND PROCESSING A PARALLEL CORPUS OF BUSINESS TEXTS FOR A MACHINE TRANSLATION SYSTEM ON NEURAL NETWORKS

The diploma paper includes 74 pages, 23 drawings, 1 table, 4 appendixes and is based on 44 sources.

Key words: MACHINE TRANSLATION, NEURAL NETWORKS, TEXT CORPUS, PARALLEL CORPUS, LINGUISTIC SOFTWARE, AUTOMATED TRANSLATION.

The objective is to identify the role of text corpora in machine translation systems based on neural networks.

The research object is machine translation systems.

The research subject is a text corpus as part of machine translation systems based on neural networks.

The relevance of the study is due to the lack of a high-quality system of Russian-Belarusian business translation, as well as the representation of the codes of the Republic of Belarus in only one of the state languages (Russian).

Methods used in writing the diploma paper: analysis, description, synthesis, comparison, structural and functional method.

The scientific novelty lies in the fact that in this work, for the first time, on the basis of practical research, the main features of the operation of Russian-Belarusian machine translation systems using a parallel corpus of texts were formulated.

The practical significance of the work consists in analyzing the Code of the Republic of Belarus on Education and the Tax Code of the Republic of Belarus, translated with the help of machine translation, identifying errors and assessing the quality of translation, in creating a parallel Russian-Belarusian text for a machine translation system on neural networks.

The research results were approved at the 77th Scientific Conference of Students and Postgraduate of BSU, Minsk, 07.04.2020; 78th Scientific Conference of Students and Postgraduate of BSU, Minsk, 05/13/2021 (presentation of the report "Creation and processing of a parallel corpus of business texts for a machine translation system on neural networks").

The author of the paper confirms that the material correctly and objectively reflects the state of the investigated process and all borrowed theoretical, methodological and methodical concepts from literature and other sources are accompanied by links to their authors.