THE UNIVERSITY of EDINBURGH

Institute for Language, Cognition and Computation
School of Informatics
The University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh EH8 9AB
United Kingdom
Tel: +44(0)131 650 4665
Fax: +44(0)131 650 6626
level3admin@inf.ed.ac.uk
www.inf.ed.ac.uk

15th September, 2022

Review of Thesis: Aidana Karibayeva

The topic of this thesis is Machine Translation (MT) and in particular MT into Kazakh from Russian and English. For cultural, scientific, commercial and other reasons we frequently need to read texts written in other languages, and if we do not speak the language, then we require the services of a translator. Since there will never be enough translators to cover the need, we seek to expand the reach of translation using automation – machine translation. For speakers of Kazakh translation is especially important because there are many more resources (e.g scientific and technical literature) available in more widely spoken languages such as English and Russian. Kazakh, however, as a Turkic language, presents specific challenges for MT largely due to the agglutinative nature of its morphology. There is also a lack of data and tools available for Kazakh translation. MT into Turkic languages has not been widely studied in the MT literature, and there are very few works considering Kazakh specifically.

This thesis seeks to improve MT for Kazakh by addressing the segmentation problem. In particular, the candidate has worked with the complete set of endings (CSE) approach to segmentation. Ms. Karibayeva extended the coverage of CSE, developed a novel automatic segmenter based on CSE, and applied the results to improve neural machine translation (NMT). The CSE-based MT system shows an improvement (on automatic scores) over a BPE-based system. BPE (byte-pair encoding) is generally held to be a hard baseline to beat. The proposed CSE-based segmentation could be applied to other Turkic languages, and segmentation is an important component in many different natural language processing (NLP) tools. An implementation of the segmentation algorithm has been released in github, as open-source software.

The results of the thesis has been published in 20 different research articles. These include 6 journal articles of which two have the candidate as the first author.

I consider that the dissertation "Development and research of models and methods of morphological segmentation of Kazakh texts for neural machine translation" submitted by Karibayeva A. satisfies the requirements for obtaining the degree of Doctor of Philosophy (Ph.D.) inthe specialty 6D070300 – Information Systems.

Best wishes

Barry Haddow
(Foreign supervisor)

informatics

/Логотип Эдинбургского университета/

ЭДИНБУРГСКИЙ УНИВЕРСИТЕТ

Институт языка, познания и вычислений
Школа информатики
Эдинбургский университет
Форум информатики
10 Crichton Street
Edinburgh EH8 9AB
Великобритания
Тел: +44(0)131 650 4665
Факс: +44(0)131 650 6626
level3admin@inf.ed.ac.uk
www.inf.ed.ac.uk

15 сентября 2022 г.

Рецензия на диссертацию: Айдана Карибаева

Тема диссертации – машинный перевод (МП) и, в частности, МП на казахский язык с русского и английского языков. По культурным, научным, коммерческим и другим причинам нам часто приходится читать тексты, написанные на других языках, а если мы не говорим на этом языке, то нам требуются услуги переводчика. Поскольку переводчиков никогда не будет достаточно для удовлетворения потребностей, мы стремимся расширить охват перевода с помощью автоматизации - машинного перевода. Для носителей казахского языка перевод особенно важен, потому что существует гораздо больше ресурсов (например, научно-технической литературы), доступных на более распространенных языках, таких как английский и русский. Казахский, однако, как тюркский язык, представляет особые проблемы для МП во многом из-за агглютинативного характера его морфологии. Также не хватает данных и инструментов для перевода на казахский язык. МП в тюркских языках не получил широкого изучения в литературе МП, и существует очень мало работ, рассматривающих именно казахский язык.

Эта диссертация направлена на улучшение МП для казахского языка путем решения проблемы сегментации. В частности, кандидат работал с методом сегментации полного набора окончаний (ПНО). Г-жа Карибаева расширила охват ПНО, разработала новый автоматический сегментатор на основе ПНО и применила результаты для улучшения нейронного машинного перевода (НМП). Система МП на основе ПНО показывает улучшение (по автоматическим оценкам) по сравнению с системой на основе КБП. КБП (кодирование байтовой пары), как правило, считается трудной базовой линией для преодоления. Предлагаемая сегментация на основе ПНО может быть применена к другим тюркским языкам, и сегментация является важным компонентом во многих различных инструментах обработки естественного языка (ОЕЯ). Реализация алгоритма сегментации была выпущена на github в виде программного обеспечения с открытым исходным кодом.

Результаты диссертации были опубликованы в 20 различных научных статьях. К ним относятся 6 журнальных статей, в двух из которых кандидат выступает в качестве первого автора.

Считаю, что диссертация «Разработка и исследование моделей и методов морфологической сегментации казахских текстов для нейронного машинного перевода», представленная Карибаевой А. удовлетворяет требованиям для получения степени доктора философии (Д.Ф.) по специальности 6D070300 – Информационные системы.

С уважением,

/подпись/

Барри Хэддоу
(Супервайзер иностранных студентов)

информатика

Перевод с английского языка на русский язык выполнен переводчиком Ахатовым Али Ахмедовичем. ИИН: 981203301255

Подпись: Ахатов Али Ахмедович

Нотариус, свидетельствуя подлинность подписи, не удостоверяет фактов, изложенных в документе, а лишь подтверждает, что подпись сделана определенным лицом.

Республика Казахстан, город Алматы.
Девятнадцатое декабря две тысячи двадцать второго года.

Я, Бекешбаева Роза Пернебековна, нотариус города Алматы, государственная лицензия №15022328 выдана Министерством Юстиции Республики Казахстан от 25.12.2015 года, свидетельствую подлинность подписи переводчика Ахатова Али Ахмедовича. Личность переводчика установлена, дееспособность и полномочия проверены.

Зарегистрировано в реестре за № **13187**
Взыскано: по ставке – 92 тенге
тех.работа – осв. на осн. п.п.1, п.2,
ст.30-1 Закона РК «О нотариате».

Нотариус: _____

ES43054412212191804 18V473301

Нотариаттық іс-әрекеттің бірегей нөмірі / Уникальный номер нотариального действия