**ОТЗЫВ**

о выпускной квалификационной работе

Якубы Надежды Михайловны

«**Оценка методов автоматического выявления устойчивых словосочетаний**»

Выпускная квалификационная работа Н. М. Якубы посвящена актуальной теме — оценке статистических мер для выявления устойчивых словосочетаний. Работа состоит из введения, четырех глав, заключения, списка литературы и четырех приложений (они занимают половину объема всей диссертации). Надежда Михайловна демонстрирует уверенное знание литературы по данному вопросу, навыки реферирования, умение формулировать и реализовывать конкретные задачи исследования.

В первой главе обсуждаются теоретические основы работы, вводятся понятия сочетаемости, свободных и несвободных словосочетаний. Автор довольно подробно излагает историю вопроса, приводит определения устойчивых словосочетаний у разных авторов.

Вторая глава посвящена методам выявления устойчивых словосочетаний. Уделяется внимание вопросам разметки текстов, лемматизации, поверхностному и глубокому синтаксическому анализу. В качестве статистического аппарата для выделения словосочетаний были выбраны и описаны 5 метрик (всего их существует более 80).

Третья глава целиком посвящена оценке методов по выделению устойчивых словосочетаний. Это известные характеристики: полнота, точность, f-мера, средняя точность. Автор справедливо замечает, что необходимо предварительно составить список словосочетаний, с которым можно сравнивать полученные результаты, — золотой стандарт. В данной работе он насчитывает 3,5 тыс. словосочетаний, взятых из разных словарей (довольно внушительный объем!).

Практическая часть работы представлена в четвертой главе. В ней описывается эксперимент по выделению биграмм в корпусе с последующим сравнением с золотым стандартом. Было найдено более 5000 именных групп, которые были упорядочены по значения каждой из 5 мер. Надеждой Михайловной были проанализированы полученные результаты и проведена оценка статистических метрик, показывающая, что maximum likelihood наиболее успешна.

При знакомстве с работой возник ряд вопросов и замечаний, которые хотелось бы прояснить.

1. Чем был продиктован выбор программы Mwetoolkit для проведения экспериментов?
2. Какова точность морфологической разметки анализатора TreeTagger? Каким образом снималась неоднозначность?
3. На чем был основан выбор данных 5 статистических мер?
4. Достаточно ли проанализировать двести словосочетаний в каждом из полученных ранжированных списков для того, чтобы сделать выводы о применимости статистических метрик?
5. Работу украсил бы перечень правил, которые были разработаны автором для нахождения словосочетаний в корпусе.

Отмеченные моменты не влияют на качество работы в целом и носят дискуссионный характер.

Работа выполнена на высоком уровне, что характеризует автора как специалиста, способного к самостоятельной научной деятельности. Выпускная квалификационная работа Н. М. Якубы представляет собой законченное исследование и отвечает всем требованиям, предъявляемым к работам подобного рода, и заслуживает высокой оценки.

Старший преподаватель

кафедры математической лингвистики СПбГУ,

к.ф.н. Хохлова М. В.

18.06.2015