

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра информационных систем управления

Аннотация к дипломной работе

**«АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ В ТЕКСТОВЫХ
ДОКУМЕНТАХ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ТИПА «ЧАСТЬ-
ЦЕЛОЕ» И «ОБЩЕЕ-ЧАСТНОЕ»**

Талдыкин Вячеслав Сергеевич

Научный руководитель – ст. преподаватель, Рубашко Н.К.

Минск, 2016

Реферат

Дипломная работа, 47 страниц, 21 рисунок, 2 таблицы, 10 источников.

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ В ТЕКСТОВЫХ ДОКУМЕНТАХ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ТИПА «ЧАСТЬ-ЦЕЛОЕ» И «ОБЩЕЕ-ЧАСТНОЕ»

Объект исследования - семантические отношения типа «часть-целое» и «общее-частное».

Цель работы - разработка алгоритма для автоматического распознавания семантических отношений «часть-целое» и «общее-частное» в текстовых документах, а также разработка программного обеспечения, которое реализует построенный алгоритм.

Методология проведения работы: системный подход, открытые системы, инженерия знаний.

Результаты работы: модели, алгоритмы, графическое приложение.

Область применения результатов: поисковые системы, базы знаний.

В данной работе исследуется и решается задача автоматического распознавания в текстовых документах семантических отношений типа «часть-целое» и «общее-частное». Определены главные проблемы и выделены основные подзадачи. Исследованы находящиеся в свободном доступе реализации лингвистических процессоров, осуществляющие синтаксический, лексический и морфологический анализы текста. Экспериментальным путём выделен лингвистический процессор Stanford parser, наиболее подходящий для решения поставленной задачи. Построен тестировочный корпус текстов, состоящий из 100 предложений, которые включают в себя примеры рассматриваемых семантических отношений. Разработана технология построения правил (паттернов), определяющих отношения «часть-целое» и «общее-частное» в предложении. Построен набор из 15-ти правил, определяющих отношение «часть-целое» и 10-ти правил, определяющих отношение «общее-частное». Разработан инструментарий, который позволяет извлечь из дерева разбора по заданным правилам пары именных групп, состоящих в отношении «часть-целое» и «общее-частное». Построена принципиальная схема решения задачи автоматического распознавания в текстовых документах отношений «часть-целое» и «общее-частное», а также разработано графическое приложение, реализующее построенное решение. Рассчитаны полнота и корректность построенного решения – 75% и 68% соответственно.

Abstract

Diploma thesis, 47 pages, 21 figures, 2 tables, 10 sources.

AUTOMATIC RECOGNITION OF SEMANTIC RELATIONS SUCH AS «PART-WHOLE» AND «COMMON-PRIVATE» IN TEXT DOCUMENTS

Object of research - semantic relationships such as "part-whole" and "common-private".

Purpose - develop an algorithm for the automatic recognition of semantic relations "part-whole" and "common-private" in text documents, as well as the development of software that implements the algorithm is built.

The methodology of the work: system approach, open systems, knowledge engineering.

Results of work: models, algorithms, graphical application.

Application of results: search engines, knowledge bases.

In this paper, we the problem of the automatic recognition of semantic relations such as "part-whole" and "common-private" in text documents is investigated and solved. During the work, the main problems have been defined and the major subtasks were identified. Publicly available implementations of the linguistic processors performing syntactic, lexical and morphological analysis of the text were studied. By experimental way language processor Stanford parser was highlighted, which is the most suitable for the task. Verifying corpus consisting of 100 proposals was built, which includes examples of considered semantic relations. The technology of construction of the rules (patterns) describing the relations "part-whole" and "common-private" in the sentence. A set of 15 rules was constructed that defines the "part-whole" relation and 10 rules that determine the "total-private" relation. Tool has been developed, which allows to extract from the parse tree for the specified rules registered couples of groups being in the "part-whole" and "common-private" relations. The principled scheme of solving the problem of automatic recognition in text documents "part-whole" and "common-private" relations was built, and developed a graphical application that implements the solution constructed. Completeness and correctness of the constructed solutions were calculated - 75% and 68%, respectively.