УДК 81–13

Оксана Игоревна Антропова,
старший преподаватель
Физико-технологический институт
Екатерина Алексеевна Огородникова,
магистрантка 2-го курса
Уральский гуманитарный институт
Уральский федеральный университет

ПРИМЕНЕНИЕ СЛОВАРНЫХ ДЕФИНИЦИЙ ПРИ АВТОМАТИЧЕСКОМ ИЗВЛЕЧЕНИИ РОДО-ВИДОВЫХ ГЛАГОЛЬНЫХ ПАР*

В статье рассматриваются различные методы автоматического извлечения семантических отношений. Обосновывается актуальность разработки подобных методов для родо-видовых связей глагольной лексики русского языка. Целью данного исследования является создание метода автоматического извлечения родо-видовых глагольных пар на материале словарных дефиниций. В ходе работы используются два основных подхода: на основе лексических маркеров и с применением синтаксических моделей.

Ключевые слова: глагол, семантика, лексикография, родо-видовые отношения, гипонимия, автоматическая обработка текста.

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-312-00129.

Oksana Antropova,
Senior Lecturer
Institute of Physics and Technology
Ural Federal University
Ekaterina Ogorodnikova,
Master student, 2 year
Ural Institute of Humanities
Ural Federal University

THE USE OF DICTIONARY DEFINITIONS BY AUTOMATIC EXTRACTION OF GENUS-SPECIES VERB PAIRS*

The article reviews various methods of automatic extraction of semantic relations. The relevance of the development of such methods for the genus-species relations of the Russian verbal vocabulary is justified. The purpose of this study is to create a method for automatic extraction of genus-species verb pairs on the material of vocabulary definitions. During the work two main approaches are used: the usage of lexical markers and syntactic models.

Keywords: verb, semantics, lexicography, genus-species relations, hyponymy, automated language processing.

Установление родо-видовых отношений между языковыми единицами — одна из главнейших задач в построении электронных тезаурусов, которые играют важную роль в автоматической обработке текстов и разнообразных лингвистических исследованиях, а также могут быть использованы в преподавании иностранного и родного языка.

Родо-видовые отношения представляют особую ценность для электронных лексикографических ресурсов, так как позволяют устанавливать логические связи общих и частных понятий. Однако проблемой остается недостаточная разработанность этой области семантики — это связано со спецификой традиционной лексикогра-

* The reported study was funded by RFBR according to the research project № 18-312-00129

фии, так как привычные бумажные словари и тезаурусы преимущественно не включают родо-видовые отношения.

В ходе работы были проанализированы основные достижения исследователей в области разработки методов ручного и автоматического извлечения семантических отношений. Так, существуют варианты установления родо-видовых отношений через перевод Princeton WordNet (данный ресурс считается наиболее полным и качественным на сегодняшний день) [1]. Возможно получение данных на основе таких баз знаний, как Wikipedia и Wiktionary [2, 3], их комбинирование с переводным способом [4]. Также исследователи предлагают метод определения семантических отношений в корпусе текстов при помощи машинного обучения [5]. Интересным и продуктивным способом является метод, предложенный в [6, 7]. Он основывается на идее лексико-синтаксических шаблонов, представляющих собой типичные языковые конструкции, которые объединяют лексические единицы, связанные теми или иными семантическими отношениями.

При всем разнообразии уже созданных и апробированных методов полной родо-видовой схемы русских глаголов не существует. Большая часть работ основывается на материале английского языка, а в русском языке базой для исследований чаще всего являются существительные.

Целью исследования является разработка метода автоматического извлечения родо-видовых глагольных отношений на основе словарных дефиниций.

На первом этапе разработки метода было доказано, что в случае с глагольной лексикой русского языка обнаружение родо-видовых отношений в корпусных данных невозможно. В качестве материала для разработки метода наиболее целесообразно использовать словарные определения, так как большинство из них строится по принципу от частного к общему, где более общая, или видовая, лексема находится в толковании для более частной, или родовой.

Вторым этапом разработки метода стало выявление группы лексических маркеров, которые сигнализируют о наличии видовой лексемы в определении. Например, такие слова, как «быстро», «медленно», «резко» часто сопровождают видовой глагол в дефиниции. Принцип обработки словарных данных был протестирован на группе

глаголов движения. Результаты, полученные при обработке дефиниций, показали достаточно высокую точность, однако метод все же представляется достаточно трудоемким, так как предполагает ручное выделение маркеров для каждой семантической группы.

Следующим шагом стало использование синтаксических анализаторов при автоматическом анализе дефиниций. Такой подход не требует ручного подбора маркеров и позволяет одинаково обрабатывать глаголы любых семантических групп. В настоящий момент ведется работа по оценке и сравнению нескольких моделей, использующих автоматическую разметку частей речи и синтаксические анализаторы.

Литература

1. *Pianta E., Bentivogli L., Christian G.* MultiWordNet. Developing an aligned multilingual database // Proc. of the 1st International WordNet Conference. 2002. January 21–25. Mysore, India. P. 293–302.

2. *Panchenko A., Adeykin S., Romanov P., Romanov A.* Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia // Concept Discovery in Unstructured Data Workshop (CDUD) of International Conference On Formal Concept Analysis. Belgium, 2012. P. 78–88.

3. *Zesch T., Müller Ch., Gurevych I.* Extracting lexical semantic knowledge from Wikipedia and Wiktionary // Proc. of the Sixth International Conference on Language Resources and Evaluatio. Marrakech, 2008. P. 1646–1652.

4. *Navigli R., Ponzetto S.* BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network // Artificial Intelligence, 193, 2012. P. 217–250.

5. *Jurafsky D., Martin J. H.* Speech and Language Processing. 2017. 499 p.

6. *Hearst M. A.* Automatic Acquisition of Hyponyms from Large Text Corpora Proc. of the 14th conference on Computational linguistics. 1992. V. 2. Nantes. P. 539–545.

7. *Hearst M. A.* Automated Discovery of WordNet Relations // ed. by C. Fellbaum. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, 1998. P. 132–152.