Зеленые страницы

УДК 81-13 DOI: 10.14529/ling190207

ВОЗМОЖНОСТИ АВТОМАТИЗИРОВАННОГО ВЫДЕЛЕНИЯ ГИПО-ГИПЕРОНИМИЧЕСКИХ ПАР ИЗ СЛОВАРНЫХ ОПРЕДЕЛЕНИЙ ГЛАГОЛОВ

О.И. Антропова, Е.А. Огородникова
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина, г. Екатеринбург, Россия

В статье рассматриваются методы автоматизированного выявления семантических отношений между единицами языка. Целью работы является разработка метода извлечения родовидовых глагольных отношений из словарных дефиниций. Предлагаемый метод основывается на идее лексико-синтаксических шаблонов и его адаптации к глагольной лексике. В ходе работы проанализированы определения из шести словарей, из дефиниций с помощью предварительно определенных лексических маркеров извлечены предполагаемые родо-видовые пары. Результат автоматической обработки проверен вручную. Проанализированы результаты применения метода, наиболее частотные ошибки, допущенные в ходе компьютерной обработки, а также представлены возможные пути повышения качества работы программы. Первичная апробация метода показала, что его дальнейшее развитие и уточнение имеет большую перспективу, и он может стать основой для создания программы автоматического установления родо-видовых отношений, показывающей высокую точность. Результаты работы могут быть использованы в различных областях прикладной лингвистики, а также при дальнейшем развитии теоретической семантики.

Ключевые слова: глагол, тропонимия, лексико-синтаксические шаблоны, гипонимия, семантика, лексикография

Современные исследования в области семантики единиц языка все больше стремятся к формализации и представлению языкового материала в унифицированном виде. Ярким примером такой тенденции являются различные идеографические словари и словари-тезаурусы, электронные ресурсы, осуществляющие автоматическую обработку текста, программы автоматического перевода и многое другое.

Семантические связи, включенные в структуру данных ресурсов, позволяют более полно и наглядно охарактеризовать каждую лексическую единицу. С помощью синонимов, антонимов, гипо-гиперонимических пар и других семантически связанных единиц человек может составить для себя представление о роли слова в системе языка, о контекстах, в которых оно может быть использовано.

Несмотря на то, что такие парадигматические отношения часто кажутся вполне естественными и очевидными, они все же являются результатом осмысления языка, то есть продуктом анализа всего лексического пространства человеком. Создание полного списка, например, слов-синонимов, возможно только при участии лингвиста-эксперта.

Учитывая количество слов в языке и количество семантических отношений (синонимия, антонимия, гипонимия, конверсивность, меронимия, отношения несовместимости и другие), ни один человек не сможет качественно обработать все единицы языка, определив все необходимые признаки, в течение целой жизни. Данная проблема находит решение в современных компьютерных технологиях. Большой популярностью пользуются автоматические методы обработки естественного языка, так как они значительно ускоряют работу, а в некоторых случаях помогают объективизировать результат, так как любой человек оценивает материал с точки зрения своего языкового опыта и субъективного видения.

На настоящий момент существует ряд разработанных методов автоматического и автоматизированного выделения разных семантических отношений. Рассмотрим некоторые из них.

Одним из вариантов построения семантической сети в любом языке является перевод отношений из Princeton WordNet на целевой язык [19]. WordNet – особый тип тезауруса, состоящий из синсетов, или синонимических рядов, которые связаны различными типами семантических отношений [11, с. 8]. Первый такой тезаурус был создан на материале английского языка в Принстонском университете, на сегодняшний день он считается наиболее полным и качественным.

Семантические отношения также могут быть извлечены из существующих баз знаний, например, Wikipedia, Wiktionary [18, 20]. В работе [17]

метод извлечения отношений из таких баз данных скомбинирован с автоматическим переводом Princeton WordNet.

Одним из возможных вариантов обнаружения семантических отношений является извлечение из корпусов текстов при помощи машинного обучения [15]. Для этого необходимы размеченные вручную данные. Их получение – это длительный и трудоемкий процесс. При этом уровень итоговой системы зависит от количества и качества исходных данных.

Работа [12] представляет оригинальный метод, разработанный на базе когнитивной теории терминологии [10]. Этот метод позволил выделить родо-видовые отношения между английскими глаголами, характерными для узкой предметной области (вулканологии) на основе автоматического построения синтаксических деревьев для корпуса научных статей по вулканологии.

Приведенные примеры методов автоматического выявления семантических отношений являются далеко не полным перечнем возможных вариантов. Все они достойны внимания, имеют свои преимущества и недостатки и могут быть в большей или меньшей степени использованы при обработке различных языков и частей речи.

В рамках данной работы стоит отдельно рассмотреть еще один метод – метод анализа языкового материала на основе лексико-синтаксических шаблонов. Такой метод впервые был сформулирован и протестирован на основе англоязычных текстов исследовательницей Марти Хэрст [13, 14], она выявила набор лексико-синтаксических шаблонов для извлечения гипонимических отношений существительных из контекстов. Например, в английском языке для этой цели могут быть использованы следующие устойчивые конструкции: such NP[1] as {NP,}*{(or | and)} NP; NP {, NP}*{,} or other NP и другие.

Подобные шаблоны могут быть обнаружены и на материале русского языка. Например, в диссертации Ю.А. Киселева [4] рассматривается алгоритм работы конструкции «такой ИГ[2] как {ИГ,}*{(и | или)} ИГ» для автоматического извлечения родо-видовых отношений существительных русского языка. Данную конструкцию можно проиллюстрировать следующим предложением: «Растет такая ягода, как клубника, малина и смородина». В этой же работе предложен метод построения лексико-синтаксических шаблонов на базе словарных определений.

Такой метод имеет свои основания, так как большая часть словарных статей составляется по определенному принципу, в котором определяемое слово чаще всего является гипонимом для главного слова в определении, и только сравнительно немногие слова толкуются через синонимы или отсылки к словам других частей речи. В работе [7] отмечается, что разработанный метод показал высокое качество результатов.

Если для существительных русского языка уже сложился определенный набор методов автоматического установления родо-видовых связей, то по отношению к глаголам подобных исследований не так много. Существующие родо-видовые глагольные схемы по разным причинам не подходят для полноценного использования в прикладных задачах, в то время как необходимость в такой классификации все возрастает. Поэтому проблема автоматического установления родо-видовых отношений между глаголами, или отношений тропонимии[3], остается актуальной.

Основываясь на идее метода лексико-синтаксических шаблонов, можно предположить, что в языке существуют определенные устоявшиеся конструкции, которые регулярно сочетают в своем составе гипероним и тропоним. Такое предположение можно проверить с помощью готового списка тропонимических пар, включающего не все глаголы русского языка, а некоторый набор, достаточный для проверки гипотезы. В статье [7] приводятся примеры таких тестовых проверок в Национальном корпусе русского языка[4], и, опираясь на результаты данных тестов, можно сделать вывод, что в русском синтаксисе не существует характерных для глаголов лексико-синтаксических шаблонов, с помощью которых возможно автоматическое извлечение тропонимических пар из корпуса.

Например, в предложении «Все *шло*, ехало, валило и *маршировало* к новому трамвайному депо, из которого ровно в час дня должен был выйти первый в Старгороде электрический трамвай» (И. Ильф, Е. Петров «Двенадцать стульев») гипероним *идти* и его тропоним *маршировать* используются в качестве контекстных синонимов. Очевидно, что при таком совместном использовании глаголов лексико-синтаксический шаблон, типичный только для отношения гипонимии, не может быть выделен [16].

Несмотря на невозможность применения метода лексико-синтаксических шаблонов к глагольной лексике русского языка в чистом виде, можно предположить, что желаемые результаты может принести автоматическая обработка лексикографических ресурсов с помощью определенных схем и правил. Данное предположение, во-первых, основано на идее, что словарные определения глаголов, так же как и существительных, чаще всего строятся на базе родо-видовых пар. Во-вторых,

[1] Noun phrase.

[2] Именная группа.

[3] Тропонимия (от греч. tropos — «способ, манера») – понятие, обозначающее родо-видовую глагольную связь, термин был впервые введён создателями Princeton WordNet [16, с. 47]. В такой терминологии родовой глагол называется гиперонимом, а видовой – тропонимом.

[4] Национальный корпус русского языка. URL: http://www.ruscorpora.ru/

формализованность использования глагольных единиц в лексикографических ресурсах намного выше, чем в свободных текстах на естественном языке.

При всех положительных сторонах разработки методов извлечения семантических отношений на базе словарей существуют и некоторые препятствия. Любой лексикографический ресурс – это в первую очередь результат человеческой работы, даже если словарь составлен экспертами высокого уровня. Это не исключает некоторой субъективности толкований и неполной формализованности структуры словарных статей, так как чаще всего над словарем трудится целый коллектив авторов, в котором каждый отвечает за свой сектор. Кроме этого, словари создаются для людей – других экспертов или рядовых пользователей, обращающихся за справочной информацией. Это накладывает определенный отпечаток на стиль текста в словарной статье: она должна быть максимально проста для понимания, но при этом в ней должны быть отражены ключевые оттенки значения. Вследствие этого в толковании слова может использоваться не просто гипероним, а, например, неоднословное выражение с более общим значением или гипероним более высокого уровня.

Например, в Словаре синонимов русского языка под редакцией Л.Г. Бабенко представлено следующее определение синонимического ряда {*вдавливать / вдавить, вжимать / вжать, вминать / вмять*}: «Оказывать / оказать на кого-либо воздействие своим весом, тяжестью, тесно прижимая к чему-либо, вводя внутрь чего-либо». Оказывать / оказать воздействие – неоднословное выражение, которое, конечно, имеет родовую связь по отношению к элементам синонимического ряда. Но проблема заключается в том, что это выражение не является устойчивым в русском языке, оно используется в качестве свободного сочетания. Обоснованность использования таких выражений в роли цельных единиц в гипо-гиперонимической цепочке – достаточно спорный вопрос. Кроме того, выделение таких неоднословных элементов автоматизированными методами влечет дополнительные трудности с технической точки зрения.

При составлении дефиниции возможно использование синонима или инфинитива, выражающего не основной семантический признак определяемого слова. Соответственно, такие случаи будут создавать определенный «мусор» при автоматической обработке. Так, глагол *прервать* в Большом толковом словаре под редакцией С.А. Кузнецова имеет следующее определение: «резко, внезапно прекратить или приостановить что-л.». *Прекратить* и *прервать* являются синонимами, глагол *приостановить* имеет более нейтральное и обобщенное значение, однако все равно может интерпретироваться как член этого же синонимического ряда, гиперонимами для которого являются глаголы *закончить*, *завершить*, *окончить*. Анализ подобных случаев затрудняется также отсутствием четких критериев, разграничивающих синонимы и гиперонимы. Во многих толкованиях гипоним и гипероним могут свободно использоваться в качестве синонимов.

Определение синонимического ряда {*нападать / напасть, бросаться / броситься, кидаться / кинуться, набрасываться / наброситься, накидываться / накинуться, обрушиваться / обрушиться, налетать / налететь, наскакивать / наскочить*} в Словаре синонимов русского языка является примером дефиниции, которая в принципе не содержит гипероним, а только инфинитив глагола, выражающего дополнительный семантический признак: «резко или неожиданно, с враждебными намерениями, начинать / начать действовать против кого-, чего-либо». Видовая пара *начинать / начать* передает значение инициации действия, но не передает наиболее общую семантику, как мог бы передать, например, глагол *двигаться / двинуться*. В данном случае он бы являлся наиболее подходящим гиперонимом.

Анализ словарных статей из различных лексикографических ресурсов показал, что выявление лексико-синтаксических шаблонов в чистом виде не дает желаемых результатов. Дело в том, что, в отличие от существительных, определения глаголов не содержат универсальных маркеров, подобных таким, как «род», «вид», «тип», «разновидность», «сорт» и т. д. Такими маркерами для глаголов могли бы быть фразы «определенным образом» или «в такой-то манере», но они не встречаются ни в одном определении глаголов из доступных нам в электронном виде словарей.

Однако наблюдается ряд других закономерностей, объединяющих дефиниции по разным признакам. Главным определяющим словом в дефиниции чаще всего является гипероним, который уточняется при помощи ряда более или менее стандартных и повторяющихся модификаторов. Например, для значительной части глаголов движения гиперонимы в определениях уточняются с помощью таких слов или выражений, как «в определенном направлении», «быстро», «медленно», «вверх», «вниз», «откуда-либо» и др. Такие **маркеры** могут варьироваться от словаря к словарю, но многие из них остаются неизменными во многих ресурсах. В этом наблюдаются определенные связи с идеей универсальных семантических единиц, или примитивов, выраженной еще А. Вежбицкой [2].

Значение семантически более сложного элемента должно толковаться через более простые единицы. Зачастую именно гипероним является более употребительным, нейтральным и «простым», а смысл более конкретного понятия можно передать, используя его гипероним и уточняющие элементы, которые также относятся к базовой лексике, чаще всего не имеют явной стилистической окраски и понятны всем носителям языка. Доста-

точно разговорный глагол *лихачить* в Большом толковом словаре под редакцией С.А. Кузнецова имеет следующее определение: «очень быстро ехать (ездить), подвергая себя опасности». Эту дефиницию можно разделить на два основных структурных элемента – гипероним *ехать / ездить* и распространяющее его наречие *быстро*. В целом именно сочетание этих двух слов и передает полное значение толкуемого слова.

Набор используемых маркеров напрямую зависит от рассматриваемой семантической группы. То есть маркеры, которые охватывают большую часть глаголов движения, могут практически отсутствовать в дефинициях глаголов мышления или социального взаимодействия. В данной работе акцент был сделан на группе глаголов движения.

Анализ словарных определений глаголов показал, что гипероним, как правило, стоит в форме инфинитива. Относительно свободный порядок слов в русском языке несколько осложняет автоматическое использование маркеров, поскольку маркер может находиться как перед определяемым инфинитивом, так и после него. Например, «*лететь / гонять / бегать / носиться / гоняться – **двигаться*** очень ***быстро*** в разных направлениях, взад и вперед (о людях и животных)» или *«лететь / мчаться / мчать / бежать / нестись – **быстро двигаться*** вперед в определенном направлении, поочередно отталкиваясь ногами от земли». Кроме того, между маркером и определяемым инфинитивом может стоять произвольное количество слов: «*влететь / взбежать / взбегать / вбегать / взлететь / вбежать / влетать / взлетать – **быстро*** или бегом ***подниматься/подняться*** куда-либо, на что-либо» или «*вносить / затаскивать / внести / втащить / затащить / втаскивать – **переносить/ перенести*** на себе кого-, что-либо ***внутрь*** чего-либо, в какое-либо помещение (часто с трудом)» (примеры из Словаря синонимов русского языка под редакцией Л.Г. Бабенко).

Тем не менее приведенные примеры иллюстрируют, что гиперонимом часто является ближайший к маркеру инфинитив. Чтобы проверить возможность применения данного наблюдения для извлечения гиперонимов из определений глаголов, была написана компьютерная программа. Сначала она извлекает определения глаголов, содержащие следующие однословные маркеры: «быстро», «медленно», «вверх», «вниз», «резко», «беспорядочно», «часто», «внутрь», «взмахом». Затем для каждого слова из отобранных на предыдущем шаге определений при помощи морфологического анализатора MyStem[5] автоматически определяет его часть речи. При этом среди всех форм глаголов особо выделяет инфинитив. Потом для каждого определения программа находит ближайший к маркеру инфинитив. Такой инфинитив программа считает искомым гиперонимом.

[5] https://tech.yandex.ru/mystem/

Данная программа была использована для анализа определений глаголов из 6 толковых словарей, указанных в таблице. Были использованы именно эти словари, поскольку авторам доступны их оцифрованные версии в формате json. Из 120 421 определений удалось выделить 1 168 определений с гиперонимами (см. таблицу).

Словарь	Всего определений глаголов	С выделенными гиперонимами	С выделенными гиперонимами, %
Бабенко [8]	1579	82	5,19
БТС [1]	17998	204	1,13
Ефремова [3]	35350	282	0,80
МАС [6]	28277	221	0,78
РуТез [5]	9114	224	2,46
Ушаков [9]	28103	155	0,55

Успешность применения данного метода зависит от характеристик конкретного словаря, от степени свободы формулировки словарного определения. Для первичного анализа результатов автоматической обработки был выбран «Словарь синонимов русского языка» под ред. Л.Г. Бабенко, так как при обработке этого словаря описанный метод дает наибольшее относительное покрытие материала.

Как видно из таблицы, этот метод обладает весьма низкой полнотой, поскольку в среднем позволяет выделить гиперонимы лишь из 1,82 % определений словаря. Точность метода, как показывает предварительный анализ, несколько выше.

Благодаря ручному анализу гиперонимов, выделенных автоматически из определений словаря под редакцией Л.Г. Бабенко, были получены следующие первичные оценки эффективности метода: среди 82 выделенных гиперонимов, проверенных вручную, 49 действительно являются таковыми. Что соответствует 40,2 % ложноположительных результатов.

При рассмотрении ложноположительных результатов был выделен ряд типичных ошибок. Например, в некоторых определениях гиперонимы представлены неоднословными выражениями. Синонимический ряд *{вдавливать / вжимать / вмять / вдавить / вминать / вжать}* имеет определение «оказывать /оказать на кого-, что-либо воздействие своим весом, тяжестью, тесно прижимая к чему-либо, вводя внутрь чего-либо». В качестве гипероним программа определяет глагол *оказывать / оказать,* который в чистом виде, конечно, не выполняет эту функцию. В данном случае гипероним будет выражен словосочетанием *оказывать / оказать воздействие.*

Еще одна типичная ошибка происходит в определениях с указанием на добровольность / недобровольность выполнения действия. Так, ряд синонимов {*бить / трепать / сотрясти / колотить / трясти / сотрясать}* имеет определение «толчками заставлять/заставить кого-, что-либо часто двигаться из стороны в сторону, совершать

колебательные движения». Программа автоматически выделяет глагол *заставлять / заставить* в качестве гиперонима. Но данная видовая пара не передает основного значения. Решением в таком случае может быть использование словосочетания *заставлять / заставить двигаться* или невозвратного глагола *двигать / двинуть*.

Синонимический ряд {*возноситься / взмывать / воспарять / вознестись / взвиться / воспарить / подниматься / взвиваться / взлететь / подняться / взмыть / взлетать / подыматься / вздыматься / взметываться / взметнуться*} в словаре определяется как «оторвавшись от земли вверх, начинать / начать движение в воздухе». В данном случае в качестве гиперонима программа выделяет глагол *начинать / начать*, однако он отражает лишь дополнительный, а не основной признак семантики приведенного синонимического ряда. Такая неточность в автоматическом анализе также является достаточно частой.

Описанные выше трудности сложно решить автоматическими методами. Так, выделение неоднословных выражений в качестве гиперонимов весьма затруднительно. Во-первых, как показывают рассмотренные выше примеры, между составными частями такого гиперонима может стоять произвольное количество любых распространяющих слов. Во-вторых, такие составные гиперонимы не являются устойчивыми выражениями. Поэтому такие случаи вряд ли возможно отличить посредством подсчета частот в корпусе от однословного гиперонима, имеющего зависимые слова. Третий тип ошибок вообще невозможно исправить автоматически, поскольку автор словарной статьи употребил слово, отражающее лишь дополнительный, а не основной признак семантики толкуемого слова. То есть толкование не содержит истинный гипероним, и при этом, насколько нам известно, нет никаких формальных признаков, позволяющих отличить истинный гипероним от слова, отражающего только дополнительный признак.

Тем не менее некоторые ошибки можно исправить автоматически. Например, для синонимического ряда {*вколотить / заколачивать / вбить / загонять / загнать / забивать / забить / вогнать / вколачивать / вгонять / заколотить / вбивать*} в определении «ударять / ударить по какому-либо предмету, заставляя его войти внутрь чего-либо» автоматически был выделен ложный гипероним *войти* и проигнорирован действительный гипероним *ударять / ударить*. Подобную и некоторые другие ошибки можно исправить автоматически, если использовать синтаксический анализатор для автоматического построения синтаксических деревьев. Тогда гиперонимом будет считаться распространенный инфинитив, находящийся в вершине синтаксического дерева. Использование синтаксических деревьев представляется наиболее перспективным способом улучшения текущих результатов, поскольку этот метод может не только исправить некоторые ошибки текущей версии программы, но и обеспечить существенно более высокую полноту покрытия определений, поскольку не будет ограничен маркерами, характерными лишь для ограниченной семантической группы глаголов. В настоящий момент ведется работа по встраиванию синтаксического анализатора в разрабатываемую программу.

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

Литература

1. Большой толковый словарь русского языка / гл. ред. С.А. Кузнецов. – СПб.: Норинт, 2000. – 1536 с.

2. Вежбицкая, А. Семантические универсалии и базисные концепты / А. Вежбицкая. – М.: Языки славянских культур, 2011. – 568 с.

3. Ефремова, Т.Ф. Новый словарь русского языка. Толково-словообразовательный / Т.Ф. Ефремова. – М.: Рус. яз., 2000. – 1233 с.

4. Киселев, Ю.А. Разработка автоматизированных методов выявления семантических отношений для электронных тезаурусов: дис. ... канд. техн. наук / Ю.А. Киселев. – Самара, 2016. – 170 с.

5. Лингвистическая онтология Тезаурус РуТез. – URL: http://www.labinform.ru/pub/ruthes/index.htm

6. Малый академический словарь: В 4 т. / под ред. А.П. Евгеньевой. – 4-е изд. – 1999. – URL: http://feb-web.ru/feb/mas/MAS-abc/default.asp

7. Огородникова, Е.А. Использование лексико-синтаксических шаблонов для формализации родо-видовых отношений в толковом словаре / Е.А. Огородникова // Евразийский гуманитарный журнал. – 2017. – № 2. – С. 20–24.

8. Словарь синонимов русского языка / под общ. ред. проф. Л.Г. Бабенко. – М.: Астрель, АСТ, 2011. – 688 с.

9. Толковый словарь русского языка: В 4 т. / под ред. Д.Н. Ушакова. – 1935–1940. – URL: http://feb-web.ru/feb/ushakov/ush-abc/default.asp

10. Benitez, F. Framing Terminology: A Process-Oriented Approach / P.F. Benitez, C.M. Linares, C.M. & M.V. Exposito // Pour une traductologie proactive – Actes. – 2005. – V. 50 (4). – URL: https://id.erudit.org/iderudit/019916ar

11. Fellbaum, Ch. WordNet – An Electronic Lexical Database / Ch. Fellbaum. – Massachusetts: MIT Press, 1998.

12. Goncharova, Yu. Specialized Corpora Processing with Automatic Extraction Tools / Yu. Goncharova, B.S. Cardenas // Procedia – Social and Behavioral Sciences. – 2013. – 95. – P. 293–297.

13. Hearst, M.A. Automated Discovery of WordNet Relations / M.A. Hearst // In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database. – MIT Press, Cambridge, 1998. – P. 132–152.

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

15. Jurafsky, D. Speech and Language Processing / D. Jurafsky, J.H. Martin. – 2017. – 499 p.

16. Miller, G. Five Papers on WordNet. Princeton University: Cognitive / G. Miller, Science Laboratory – 1993. – URL: http:// wordnetcode. princeton.edu/5papers.pdf

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

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

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

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

Антропова Оксана Игоревна, старший преподаватель кафедры технической физики, Уральский федеральный университет, choksy@mail.ru

Огородникова Екатерина Алексеевна, ассистент кафедры лингвистики и профессиональной коммуникации на иностранных языках, Уральский федеральный университет, kruglikova.katya@yandex.ru

Поступила в редакцию 22 января 2019 г.

DOI: 10.14529/ling190207

POSSIBILITIES OF COMPUTER-AIDED EXTRACTION OF HYPER-HYPONYMIC PAIRS FROM DICTIONARY DEFINITIONS OF VERBS

O.I. Antropova, *choksy@mail.ru*
E.A. Ogorodnikova, *kruglikova.katya@yandex.ru*
Ural Federal University named after B.N. Yeltsin, Ekaterinburg, Russian Federation

The paper concerns computer-aided methods for extraction of semantic relations between lexical units. The aim of the research is to elaborate a method of genus-species relations extraction from dictionary definitions. The suggested method is based on the idea of lexico-syntactic patterns and its adaptation to verbal vocabulary. Definitions from six dictionaries are analyzed, probable genus-species pairs are extracted from the definitions with the help of previously defined lexical markers. The result of automatic processing is verified manually. The results of the method application and the most frequent mistakes made during computer processing are analyzed and possible ways of method improvement are suggested. The primary approbation of the method showed that its further development and elaboration have good prospects, and it can be a base for creation of an automatic genus-species relations extraction software showing high accuracy. The results of the research can be used in different domains of computational linguistics and during further research in theoretical semantics.

Keywords: verb, troponymy, lexico-syntactic patterns, hyponymy, semantics, lexicography.

References

1. *Bolshoy tolkovyy slovar russkogo yazyka*. [The Big Explanatory Dictionary of the Russian Language. Ed. S.A. Kuznetsov]. St-Petersburg, 2000, 1536 p.

2. Vezhbitskaya A. *Semanticheskiye universalii i bazisnyye kontsepty* [Semantic Universals and Basic Concepts]. Moscow, 2011, 568 p.

3. Efremova T.F. *Novyy slovar russkogo yazyka. Tolkovo-slovoobrazovatelniy* [The New Dictionary of Russian Language. Explanatory-derivational]. Moscow, 2000. 1233 p.

4. Kiselev Yu.A. *Razrabotka avtomatizirovannykh metodov vyyavleniya semanticheskikh otnosheniy dlya elektronnykh tezaurusov: dis. ... kand. tekhn. nauk* [Elaboration of Computer-aided Methods of Semantic Relations Extraction for Electronic Thesauri. Thesis]. Samara, 2016.

5. *Lingvisticheskaya ontologiya Tezaurus RuTez* [Linguistic Ontology Thesaurus RuThes]. URL: http://www.labinform.ru/pub/ruthes/index.htm (accessed: 20.12.2018).

6. Evgenyeva A.P. *Malyy akademicheskiy slovar*: *V 4 t. 4 izd.* [The Small Academic Dictionary: in 4 v. The 4th ed.]. 1999. URL: http://feb-web.ru/feb/mas/MAS-abc/default.asp (accessed: 25.12.2018).

7. Ogorodnikova E.A. *Ispolzovaniye leksiko-sintaksicheskikh shablonov dlya formalizatsii rodo-vidovykh otnosheniy v tolkovom slovare* [The Usage of Lexico-syntactic Patterns for Genus-species Relations in an Explanatory Dictionary]. *Evraziyskiy gumanitarnyy zhurnal* [Eurasian Humanitarian Journal], 2017, no. 2, pp. 20–24.

8. Babenko L.G. *Slovar sinonimov russkogo yazyka* [The Dictionary of Synonyms of the Russian Language]. Moscow, 2011. 688 p.

9. Ushakov D.N. *Tolkovyy slovar russkogo yazyka: V 4 t.* [The Explanatory Dictionary of the Russian Language: in 4 v.]. 1935-1940. URL: http://feb-web.ru/feb/ushakov/ush-abc/default.asp (accessed: 23.12.2018).

10. Benitez F., Exposito C.M., Exposito M.V., Linares C.M. *Framing Terminology: A Process-Oriented Approach. Pour une traductologie proactive, Actes,* 50 (4), 2005. URL: https:// id.erudit.org/iderudit/019916ar (accessed: 23.12.2018).

11. Fellbaum Ch. *WordNet – An Electronic Lexical Database*. MIT Press, Massachusetts, 1998.

12. Goncharova Yu., Cardenas B.S. Specialized Corpora Processing with Automatic Extraction Tools. *Procedia – Social and Behavioral Sciences*, 2013, 95, pp. 293–297.

13. Hearst M. A. *Automated Discovery of WordNet Relations.* In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, MIT Press, Cambridge, 1998, pp. 132–152.

14. Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora. *Proc. of the 14th conference on Computational linguistics, v. 2,* Nantes, France, 1992, pp 539–545.

15. Jurafsky D., Martin J.H. *Speech and Language Processing*. Stanford University, 2017, 499 p.

16. Miller G. *Five Papers on WordNet*. Princeton University: Cognitive. Science Laboratory, 1993. URL: http://wordnetcode.princeton.edu/5papers.pdf (accessed: 22.12.2018).

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

18. 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, pp. 78–88.

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

20. 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, Morocco*, 2008, pp. 1646–1652.

Oksana I. Antropova, Senior Lecturer of the Chair of Technical Physics, Ural Federal University named after B.N. Yeltsin (Ekaterinburg), choksy@mail.ru

Ekaterina A. Ogorodnikova, Teaching Assistant of the Chair of Linguistics and Professional Communication on Foreign Languages, Ural Federal University named after B.N. Yeltsin (Ekaterinburg), kruglikova.katya@yandex.ru

Received 22 January 2019

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Антропова, О.И. Возможности автоматизированного выделения гипо-гиперонимических пар из словарных определений глаголов / О.И. Антропова, Е.А. Огородникова // Вестник ЮУрГУ. Серия «Лингвистика». – 2019. – Т. 16, № 2. – С. 51–57. DOI: 10.14529/ling190207

FOR CITATION

Antropova O.I., Ogorodnikova E.A. Possibilities of Computer-Aided Extraction of Hyper-Hyponymic Pairs from Dictionary Definitions of verbs. *Bulletin of the South Ural State University. Ser. Linguistics*. 2019, vol. 16, no. 2, pp. 51–57. (in Russ.). DOI: 10.14529/ling190207