Об автоматической рубрикации терминов тезауруса открытой информационно-аналитической системы

© Бойков В.Н. © Захаров В.Е.
Институт космических исследований РАН
Москва
boykov_bh@bk.ru vezakhar@mx.iki.rssi.ru

© Каряева М.С. © Соколов В.А.
Ярославский государственный университет
Ярославль
mari.s.ka@mail.ru valery-sokolov@yandex.ru

Аннотация

В работе рассматривается применение методов лингвистического анализа для автоматического рубрицирования терминов открытого сетевого ресурса «Тезаурус по поэтологии» в составе Информационно-аналитической системы русской поэзии. Приведены основные принципы и процедуры автоматической рубрикации корпуса терминов тезауруса.

Работа поддержана Российским фондом фундаментальных исследований, грант № 13-06-00448.

1 Введение

Автоматизация семантического анализа полнотекстовой информации для извлечения релевантных данных является актуальной задачей в инженерии знаний. Это важно и для автоматического построения таких метаописаний и семантических структур предметной области, как тезаурусы и онтологии, где описываются основные понятия и отношения между ними.

Семантическая модель предметно-ориентированного тезауруса предопределяет структуру его данных, тогда как его структурирование осуществляется по мере непосредственного описания его понятий и отношений между ними. Определяющими структуру данных являются иерархические отношения, поскольку задают сложность структуры, как по числу иерархических уровней (рубрик), так и по числу вариаций (подрубрик) на одном уровне. Среди иерархических отношений чаще всего выделяются отношение «рода и видов», задающее классификацию понятий, и отношение «целого и частей», систематизирующее данные.

Задача автоматизации описания понятия, основу которого составляет его определение, решается с помощью известных методов полнотекстового поиска из авторитетных источников (баз знаний, справочников, словарей, энциклопедий).

Задача автоматического выявления отношений между понятиями сводится к извлечению этих отношений из описания понятий и требует специальных лингвистических методов, поскольку чаще всего формально эти отношения в описании не задаются. Для выявления таксономических отношений (синонимии и гиперонимии) между понятиями оказывается эффективным формирование лингвистических шаблонов [1]. Растущий интерес исследователей к лингвистическим методам анализа текста для построения онтологий связан с повышением качества синтаксических анализаторов [2].

Семантическая неопределенность сложных синтаксических конструкций заставляет обратиться к такому фундаментальному семантико-синтаксическому понятию, как синтаксема, которая является и минимальной синтаксической единицей, и носителем элементарного смысла [3]. Синтаксемы могут происходить от различных частей речи, но преимущественно – от имен существительных и представляют собой падежные или предложно-падежные словоформы в синтаксическом контексте. К сожалению, конструкции со сложными предлогами в репертуар синтаксем пока не вошли.

Примером конструктивного использования синтаксемы может служить лингвистический подход, осуществленный в [4] при построении онтологии конкретной предметной области.

2 Описание объекта структурирования

Представление о составе Информационно-аналитической системы русской поэзии (ИАСРП), а также о методологических концепциях тезауруса по поэтологии (ТП) дается в работах [5–8].

В исходной комплектации ТП содержится:

– Базовый корпус около 2000 терминов по поэтологии;

– Формуляр терминологической статьи тезауруса (ТСТ), представляющий семантическую модель ТП [8], где заданы 27 полей ТСТ трех типов – поля,

Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.

относящиеся к термину (основные – «определение» и «рубрика»), и поля иерархических и неиерархических отношений между терминами;

– Базовый рубрикатор терминов, в котором представлена экспертная рубрикация предметной области по 10 подобластям (рубрикам верхнего уровня), соответствующим дисциплинам поэтологии [5], что необходимо и достаточно для автоматической рубрикации всего корпуса терминов:

1. Стиховедение;
2. Стилистика;
3. Поэтика;
4. Риторика;
5. История литературы;
6. Переводоведение и литературная компаративистика;
7. Текстология;
8. Герменевтика;
9. Теоретические школы и направления;
10. Логика и методология науки.

В качестве примера произведена ручная рубрикация 32 терминов 4 верхних уровней подобласти 1. Стиховедение:

1.1. Стих: (кластер);
1.1.1 Метрика: (кластер);
1.1.1.1. Квантитативная метрика: (кластер);
1.1.1.2. Квалитативная метрика: (кластер);
1.1.2. Явления начала и конца стихотворной строки: (кластер);
1.1.2.1. Анакруза, анакруса: (кластер);
1.1.2.2. Каталектика: (кластер);
1.1.3. Ритмика: (кластер);
1.1.3.1. Акцентуация: (кластер);
1.1.3.2. Цезура и Словоразделы: (кластер);
1.1.4. Строфика: (кластер);
1.1.4.1. Строфы: (кластер);
1.1.4.2. Квазистрофические формы и Гиперстрофические формы: (кластер);
1.1.4.3. Твёрдые формы стиха: (кластер);
1.1.5. Рифмика: (кластер);
1.1.5.1. Типы рифмы по количеству слогов: (кластер);
1.1.5.2. Типы рифмы по фонетическому составу: (кластер);
1.1.5.3. Типы рифмы по лексическому составу: (кластер);
1.1.5.4. Рифменные последовательности: (кластер);
1.1.5.5. Квази-рифмические способы организации стиха: (кластер);
1.1.6. Лингвистика стиха: (кластер);
1.1.6.1. Звуковая организация стиха: (кластер);
1.1.6.2. Графическая организация стиха: (кластер);
1.1.6.3. Ритмико-фонетические явления в стихе: (кластер);
1.1.6.4. Морфология стиха: (кластер);
1.1.6.5. Синтаксис стиха: (кластер);
1.1.6.6. Мелодика стиха: (кластер);
1.1.6.7. Поэтическая семантика: (кластер);
1.2. Проза (в отличие от стиха): (кластер);
1.2.1. Формы прозы: (кластер);
1.2.2. Членение прозы: (кластер).

Рубрики 4 верхних уровней всех 10 подобластей представляют всего 115 терминов из двухтысячного корпуса.

ТП разрабатывается с применением Wiki-технологии. Базы знаний с использованием Wiki-технологий имеют ряд преимуществ, так как позволяют энтузиастам-исследователям самим через веб-интерфейс активно включиться в процесс редактирования базы знаний: исправления ошибок, добавления новых материалов и т.д. Коллективное редактирование ТП может ускорить наполнение ТСТ и не должно отразиться на его качестве, поскольку добавление новой информации в ТП отслеживает наряду с администратором сайта модератор системы – квалифицированный специалист в области поэтологии, который принимает или отвергает внесение или изменение контента в ТП.

Вместе с тем, несмотря на возможность получения высокого качества при ручном заполнении ТСТ, трудоемкость и множественность звеньев процесса не обеспечивает его должной скорости, поэтому задача его автоматизации на предварительном этапе представляется достаточно важной. Такая автоматизация позволяет осуществить структурирование предметной области и вследствие чего дает возможность энтузиастам-исследователям завершить описание термина в контексте его места в общей структуре ТП.

ИАСРП в своем составе должен содержать помимо ТП также аналитический блок [6], который предназначен для автоматического решения различных задач стиховедения в отношении поэтических текстов. Для постановки и алгоритмизации этих задач необходим завершенный в достаточной полноте тезаурус, что предполагает, в том числе, и его рубрикацию. В этом контексте очевидна актуальность создания программно-алгоритмического модуля для решения комплекса задач, связанных со структуризацией ТП и рубрикации его терминов.

3 Модуль автоматического структурирования ТП

Конечной целью автоматического структурирования ТП является рубрикация его терминов, т.е. отнесение каждого термина к его рубрике в иерархическом дереве, точнее, к цифровому коду его рубрики, идентифицирующему место термина в иерархии. В данном случае каждому термину определяется место в цепочке терминов, привязанной к одной из вершин базового рубрикатора терминов.

В модуле автоматического структурирования ТП выделяются следующие подмодули последовательных автоматических процедур.

Подмодуль 1: автоматическое заполнение поля ТСТ «определение»;

Подмодуль 2: автоматическое заполнение полей ТСТ «родовое понятие» и «видовые понятия»;

Подмодуль 3: автоматическое заполнение полей ТСТ «целое» и «части»;

Подмодуль 4: автоматическое заполнение полей «рубрика» и «дисциплина (рубрика первого уровня)».

Процедуры подмодуля 1

Для автоматического заполнения поля ТСТ «определение» используются следующие (в порядке репрезентативности) оцифрованные источники:

– Краткая литературная энциклопедия: В 9 т. – М.: Сов. энцикл., 1962-1978 [9].

– Квятковский А.П. Поэтический словарь. – М.: Советская энциклопедия, 1966 [10].

– Литературная энциклопедия: В 11 т. – М.: Ком. акад., 1929-1939 [11].

Дополнительно полезны также некоторые другие энциклопедии и словари [12–16].

Ключом для извлечения определения понятия из источника служит термин из имеющегося терминологического словника ТП. Все извлеченные определения для данного термина помещаются в поле ТСТ «альтернативные определения».

Затем производится лингвистическая (частеречная и синтаксическая) разметка текстов определений с помощью синтаксического анализатора (парсера), размещенного на электронном ресурсе «Автоматическая обработка текста» [17]. При разметке текста определения термина в нем отмечаются другие термины, включенные в терминологический словник, что важно для результативности процедур последующих подмодулей.

Среди альтернативных определений могут оказаться синонимические определения, а также противоречащие друг другу и нечеткие определения. Решение о помещении того или иного определения в поле ТСТ «определение» и сохранении его в поле «альтернативные определения» принимается модератором системы.

Процедуры подмодуля 2 и 3

Хотя для рубрикации достаточно заполнить поле «родовое понятие», но иногда род термина определяется только через его представление в качестве вида другого.

Первой процедурой выявления рода для данного термина является его поиск в полях «видовые понятия» в соответствующих полях других терминов: его нахождение в поле «видовые понятия» некоторого термина означает, что последний и является родом для данного.

Аналогичной процедурой выявления вида для данного термина является его поиск в полях «родовое понятие» ТСТ других терминов: его нахождение в поле «родовое понятие» некоторого термина означает, что последний является видом для данного.

Автоматизация выделения из определения термина его рода и видов с помощью лингвистических методов исходит из выявления синтаксических конструкций, задающих отношения рода и вида. Элементарные единицы русского синтаксиса (синтаксемы) для этих отношений приводятся в [3], хотя они не исчерпывают всех синтаксических конструкций для этих целей. Примеры синтаксем, несущих отношения «род-виды»:

– предмет среди класса предметов – **предлог «среди»** + **род. падеж** (выделяться, находиться среди …);

– отнесение вида к роду – **предлог «к»** + **дат. падеж** (относиться, принадлежать к …).

Для выявления «родо-видовых» отношений служат и другие синтаксические конструкции [18].

Для выявления вида:

– сложное слово, часть которого (производящая основа или словообразующая морфема) задает единство принадлежности к роду, как, например, в «метрике» различают явления «монометрии» и «полиметрии».

Для выявления как рода, так и вида:

– словосочетание, представляющее собой видовой термин, где «в качестве опорного терминоэлемента выступает родовой термин», как, например, в роду «ямб» выделяются двустопный, трехстопный, 4-стопный, 5-стопный и 6-стопный ямбы.

Для выявления «родо-видовых» отношений полезны также конструкции с предметно определенным обобщающим словом или словосочетанием при однородных членах предложения.

После выявления «родо-видовых» отношений данного термина определяющая эти отношения синтаксическая конструкция добавляется в набор шаблонов для синтаксического анализа последующих терминов.

При успешном выявлении рода или вида для данного термина может оказаться так, что соответствующих им терминов в терминологическом словнике нет, и тогда решение о внесении этих терминов в словник принимает модератор системы.

С другой стороны, в самих определениях терминов могут не найтись отсылки к роду и видам (у конечных терминов виды отсутствуют), и, следовательно, не всегда можно вывести родо-видовые цепочки к 4 верхним уровням, имеющим коды рубрик. В этом случае придется использовать открытость системы и компетентность энтузиастов-исследователей предметной области.

Процедуры выявления отношений «целое-части» для данного термина осуществляются по аналогии с предыдущими.

Специфика выделения из определения термина отношений «целое-части» с помощью лингвистических методов отличается более широким набором синтаксем, используемых для выявления этих отношений:

– обозначение частей целого – **предлог «из»** + **род. падеж** (состоять, слагаться, складываться, составляться, собираться или образовываться из …);

– часть, отделенная от целого – **предлог «от»** + **род. падеж**;

– дополнение части к целому – **предлог «к»** + **дат. падеж** (приобщенное к чему-то);

– соединение частей в целое – **предлог «в»** + **вин. падеж** (складываться, собираться в …);

– распадение целого на части – **предлог «в»** + **вин. падеж** (распадаться в …);

– деление целого на части (несколько частей) – **предлог «на»** + **вин. падеж**.

Процедуры подмодуля 4

Нахождение данного термина в одном из кластеров рубрики верхнего уровня определяет процедуру заполнения поля «дисциплина» в его ТСТ.

После выявления «родо-видовых» отношений данного термина его «родовое понятие» сверяется с терминами БРТ, имеющими код рубрики, и при его совпадении с одним из таких терминов БРТ, производится рубрикация данного термина и его перевод из кластера в рубрикатор: данному термину присваивается видовой код рубрики найденного термина. Затем код рубрики данного термина вносится в поле «рубрика» его ТСТ.

Далее производится рубрикация «видовых понятий» данного термина и их перевод из кластера в рубрикатор: им присваиваются видовые коды рубрики данного термина. Затем заполняются поля ТСТ видовых понятий данного термина «рубрика» и «родовое понятие», куда вносится данный термин.

4 Реализация автоматического заполнения поля ТСТ «определение»

На рисунке 1 для более наглядного представления размещены первые 12 полей ТСТ для конкретного термина. Следует отметить, что заполнение ТП является не только трудоемким процессом, так как необходимо заполнить порядка 50 тысяч полей, но и требует достаточного уровня знаний предметной области.

На рисунке 2 показана схема разметки ТСТ источника, где ее текст открывается термином, который содержится в терминологическом словнике и, следовательно, в БРТ, что позволяет видеть, к какой рубрике верхнего уровня относится данный термин. Соответственно, в ТП автоматически создается новая ТСТ с данным термином и заполняются поля «термин» и «дисциплина».

Далее существует 2 варианта объяснения термина: первый вариант содержит иностранный эквивалент с указанием языка и перевода термина, второй вариант встречается при условии русского происхождения термина или утраты его иностранного происхождения.

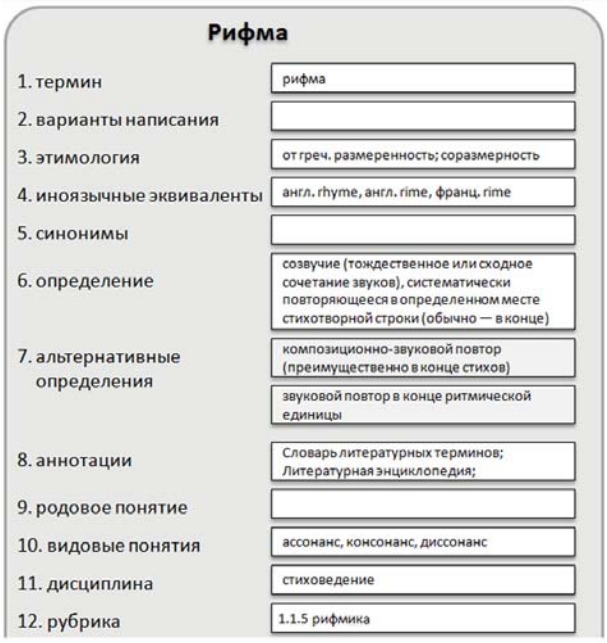

Рис. 1. Пример термина «Рифма» с заполненными полями

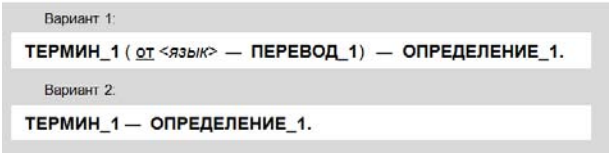

Рис. 2. Разметка «термин-определение».

Так как одно из полей тезауруса содержит поле «иноязычные эквиваленты», то использование конструкции, представленной на рисунке 2, снимает вопрос о заполнении этого поля. Разметка *<язык>* представляется в тезаурусе как дополнительный технических словарь, содержащий набор различных языков в виде «англ.», «нем.», «греч.», «франц.», «араб.» и т.д. Ключевое слово и соответствующая разметка «от» + *<язык>* автоматически указывают на конструкцию строки, которая может быть использована в качестве описания поля «иноязычные эквиваленты».

При использовании варианта 2 разметки терминологической статьи источника поле «иноязычные эквиваленты» остается пустым. Далее, для всех вариантов выделяется оставшаяся часть предложения и оформляется как определение термина.

Пример разметки статьи источника [8] и выделения из нее определения термина «рифма»:

Рифма (от греч. — соразмерность) — композиционно-звуковой повтор преимущественно в конце двух или нескольких стихов, чаще — начиная с последнего ударного слога в рифмуемых словах. В русских пиитиках (10—18 вв.) этот повтор назывался «краесогласием».

ТЕРМИН_1 := рифма

ПЕРЕВОД_1 := соразмерность

ОПРЕДЕЛЕНИЕ_1 := композиционно-звуковой повтор преимущественно в конце двух или нескольких стихов, чаще — начиная с последнего ударного слога в рифмуемых словах.

Каждый из оцифрованных источников, перечисленных выше [9–11], содержит описания и определения не более одной тысячи терминов. Безусловно, литературоведческих справочников недостаточно для полного покрытия предметной области, так как представленные выше источники содержат целый ряд идентичных терминов, что уменьшает размерность общего набора различных терминов. Поэтому неизбежно использование альтернативных и универсальных источников знаний, таких как словари и энциклопедии общей направленности.

5 Статусы терминов после автоматического структурирования ТП

После применения процедур автоматической рубрикации все термины ТП условно можно разделить на три группы:

1) Термины с заполненными полями ТСТ, в том числе с полями, определяющими отношения между терминами.

2) Термины с заполненными полями ТСТ, кроме полей, определяющих отношения между терминами.

3) Термины, которые не встретились в литературоведческих и стиховедческих источниках, соответственно, не имеют автоматически заполненных полей ТСТ.

Первая группа терминов является завершенной и имеет статус «Завершено», вторая группа терминов получает статус «В работе» и доступна для дальнейшей автоматической рубрикации, пока не приобретет статус «Завершено». Третья группа терминов имеет статус «Не определено», который показывает, что термин не подвергался автоматической рубрикации ввиду отсутствия соответствующего термина в оцифрованных источниках. Определение этих статусов указывает, что следует прибегнуть к ручному заполнению ТСТ для терминов со статусом «Не определено».

Основные три указанных источника терминов [9–11] содержат порядка 1000 терминов каждый и, если учесть далеко не полное пересечение этих совокупностей терминов, то в целом корпус терминов со статусами «Завершено» и «В работе» может составить около 1500 единиц.

Данный результат был получен с помощью статистических расчетов, проведенных вручную, что гарантирует качество проводимых исследований и может быть достоверной верхней оценкой для автоматического подхода. Экспертная ручная оценка обеспечивает возможность получения близкого результата, поскольку используемые правила могут быть алгоритмизированы. Это дает основание считать, что метод может быть автоматизирован на практике и в дальнейшем улучшен с помощью обучающих методов (в том числе, используя машинное обучение).

6 Заключение

В работе представлены методы и процедуры, которые позволяют автоматически структурировать термины такой предметной области, как «поэтология». Благодаря этому для заполнения полей ТСТ не требуется помощь специалиста, необходимо лишь реализовать алгоритмы, использующие оцифрованные литературоведческие и стиховедческие источники. В дальнейшем, при совершенствовании лингвистических методов анализа, детально рассматривающих частные случаи и исключения и использующих наряду со справочными источниками литературоведческие исследования, можно автоматизировать заполнение полей ТСТ и подвергнуть автоматической рубрикации значительную часть группы терминов, имеющих статус «Не определено».

Таким образом, имея достаточный набор неструктурированных терминов, источники знаний и ряд аналитически полученных правил, можно осуществлять автоматическую рубрикацию терминов. Кроме того, такой подход к структуризации предметной области может использоваться в более широком аспекте гуманитарного знания.

Литература

[1] M.A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora // Proceedings of the 14th International Conference on Computational Linguistics. – 1992. – P 539–545.

[2] J. Makki, A.-M. Alquier, V. Prince. Semi Automatic Ontology Instantiation in the domain of Risk Management // IFIP, Advances in Information and Communication Technology. – 2008. – Vol. 288. – P. 254–265.

[3] Г.А. Золотова. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. – М.: Наука, 1988.

[4] Е.А. Оробинская. Метод автоматического построения онтологии предметной области на основе анализа лингвистических характеристик текстового корпуса // Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012). – СПб, 2012.

[5] В.Н. Бойков, В.Е. Захаров, И.А. Пильщиков, Т.М. Сысоев. Тезаурус как инструмент поэтологии // Моделирование и анализ информационных систем. – 2010. – Т. 17, № 1. – С. 5–24.

[6] V.N. Boikov, V.E. Zakharov, M.S. Karyaeva, V.A. Sokolov. Предметно-ориентированный тезаурус в открытой информационно-аналитической системе (Domain-Specific Thesaurus as a Part of an Information-Analytical System) – RCDL-2013.

[7] В.Н. Бойков, В.Е. Захаров, М.С. Каряева, В.А. Соколов. Тезаурус по поэтологии как инструмент для информационного поиска и коллекции знаний // Моделирование и анализ информационных систем. – 2013. – Т. 20, № 4. – С. 125–135.

[8] Бойков В.Н., Пильщиков И.А. Семантическая модель «Тезауруса по поэтологии» в составе информационно-аналитической системы // Интернет и современное общество: сборник научных статей. Труды XVI Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2013). — СПб.: НИУ ИТМО, 2013.

[9] Краткая литературная энциклопедия: в 9 т. – М.: Сов. энцикл., 1962–1978. (http://feb-web.ru/feb/kle/default.asp?/feb/kle/kle.html)

[10] А.П. Квятковский. Поэтический словарь. – М.: Советская энциклопедия, 1966. (wikilivres.ru); (feb-web.ru/feb/kps/kps-abc)

[11] Литературная энциклопедия: в 11 т. – М.: Ком. акад., 1929–1939. (http://feb-web.ru/feb/litenc/encyclop/)

[12] Литературная энциклопедия. Словарь литературных терминов: в 2 т. – М., Л.: Изд-во Л.Д. Френкель, 1925. (enc-dic.com/lit)

[13] Большая советская энциклопедия: в 30 т. – 3-е изд. – М.: Сов. энцикл., 1969–1978. (http://slovari.yandex.ru/dict/bse/)

[14] Лингвистический энциклопедический словарь. М.: Советская энциклопедия, 1990. (www.tapemark.narod.ru/les)

[15] Ахманова О. С. Словарь лингвистических терминов. – М.: Сов. энцикл., 1966.

[16] Розенталь Д. Э., Теленкова М. А. Словарь-справочник лингвистических терминов. – Изд. 2-е. — М.: Просвещение, 1976. (http://www.intruderalarms.sebastopol.ua.); (http://www.gumer.info/bibliotek_Buks/Linguist/DicTermin/index.php)

[17] Автоматическая обработка текста. [Электронный ресурс] // Режим доступа: http://aot.ru/demo/morph.html

[18] Н.А. Гурдаева. Принципы структурной организации лексических терминов как результат родо-видовых отношений системы понятий // Вестник ТГПИ. Специальный выпуск 1. Таганрог, 2011.

On the Automatic Structuring of the Thesaurus for an Open Information-Analytical System

V.N. Boikov, V.E. Zakharov,
M.S. Karyaeva, V.A. Sokolov

In the work methods of using the linguistic analysis for the automatic structuring of the open network resource “Information-Analytical System of Russian Poetry” are considered. The basic principles that allow to realize a way of the automatic categorization of the thesaurus are given.