УДК 004.853

Н. И. Гурин, Я. А. Жук
Белорусский государственный технологический университет

АЛГОРИТМ ПОДГОТОВКИ ТЕКСТА ОБУЧАЮЩЕЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ К СЕМАНТИЧЕСКОМУ АНАЛИЗУ

Статья посвящена описанию алгоритма подготовки текста информационной системы к семантическому анализу. Основными блоками данного алгоритма являются морфологический анализ слов, синтаксический анализ предложений и формирование базы типов семантических связей. С целью оптимизации в алгоритме предусмотрено параллельное выполнение процессов синтаксического анализа и формирования запроса на вставку типов семантических связей, полученных на основании морфологического анализа, в таблицу реляционной базы данных. Результатами алгоритма являются база типов связей дуг и набор подготовленных к семантическому анализу предложений. Морфологический анализ необходим как для составления базы типов связей семантической сети, так и для синтаксического анализа. В рамках синтаксического анализа предложений предлагается преобразовать текст в соответствии с рядом правил. Первое правило состоит в удалении предложений и оборотов, не несущих смысловой нагрузки. Второе – в разрешении анафор в исходном тексте, т. е. замене местоимений на обозначаемые ими информационные единицы. Третье – в преобразовании сложносочиненных предложений и предложений с однородными сказуемыми в набор самостоятельных простых предложений. Выполнение данных правил обеспечивает эффективность выявления связей семантической сети информационной системы.

Ключевые слова: морфологический анализ, синтаксический анализ, обработка естественного языка, семантические сети.

N. I. Gurin, Ya. A. Zhuk
Belarusian State Technological University

ALGORITHM OF THE PREPAIRING E-LEARNING SYSTEM FOR SEMANTIC ANALYSIS

The article describes the algorithm for preparing text information system for semantic analysis. The main blocks of this algorithm are morphological analysis of words, parsing sentences and building a base of semantic relations types. The algorithm includes parallel execution of the processes of parsing sentences and the formation of the insert request for semantic relations types, obtained on the basis of morphological analysis, into a table of a relational database. Results of the algorithm are the database of semantic relationships types and the set of prepared for the semantic analysis sentences. Morphological analysis is required both for generating semantic relationships types database and for parsing sentences. The base of the parsing sentences consists of three rules. The first rule is to remove parts of sentences, that does not bear semantic load. The second rule is to resolve anaphor in the source text, i. e. make the replacement of pronouns on the referred information items. The third rule is to convert compound sentences and sentences with homogeneous predicates in the set of independent simple sentences. These rules provide the further identification of semantic network relations.

Key words: morphological analysis, syntax analysis, natural language processing, semantic networks.

Введение. Создание диалоговой информационной системы, способной точно отвечать на поставленные вопросы одиночными предложениями, требует составления базы знаний, содержащей понятия предметной области и связи между ними. Качественное извлечение знаний из текстов информационных систем остается актуальной задачей для разработчиков интеллектуальных систем. Общими чертами существующих алгоритмов извлечения знаний является выделение этапов морфологического, синтаксического и семантического анализа [1]. Наличие подготовительных этапов морфологического и синтаксического анализа обусловлено неспособностью алгоритмов семантического анализа обрабатывать предложения произвольной структуры и необходимостью предварительной подготовки базы типов семантических связей, которая будет использоваться при выявлении новых связей семантической сети.

Существующие программные средства морфологического анализа, построенные на анализе псевдоокончаний слов, позволяют достаточно точно определить морфологические признаки слов [2]. В связи с фактически произвольным порядком слов в предложениях на

русском языке, синтаксический анализ предложений является более сложной задачей [3]. Однако, в рамках русского языка существует научный функциональный стиль, который отличается подчеркнутой логичностью и точностью, достигаемыми путем использования четкой структуры предложений и специальной терминологии. Учет данных особенностей стиля позволил определить ключевые правила для подготовки текста обучающей информационной системы, составленного в научном функциональном стиле, к преобразованию в семантическую сеть диалогового модуля.

Основная часть. Для выполнения задач подготовки текста к семантическому анализу был разработан соответствующий алгоритм. Он включает в себя блоки морфологического анализа, формирования запроса на вставку типов связей в таблицу реляционной базы данных и преобразования сложных предложений, характерных для научного функционального стиля, в простые предложения, над которыми можно будет выполнить операцию актуального членения предложения в ходе дальнейшего семантического анализа. Следует отметить возможность параллельного выполнения процессов формирования запроса на вставку типов семантических связей в базу данных и подготовки текста к актуальному членению предложений. Блок-схема алгоритма подготовки текста к семантическому анализу показана на рис. 1.

Нами обнаружена в текстах научного стиля высокая встречаемость особой структуры предложений, используемая с целью обеспечения недвусмысленности высказываний. Данная структура состоит из трех частей: подлежащего с относящимися к нему дополнениями и определениями, сказуемого и других дополнительных членов предложения, связанных со сказуемым, а не с подлежащим. Важно отметить, что подлежащее с набором связанных с ним вспомогательных членов предложения в тексте научного стиля образует единый термин. Таким образом, сказуемое в таких предложениях находится между двумя информационными единицами, что позволяет автоматизировать выявление смысловых единиц после нахождения сказуемых.

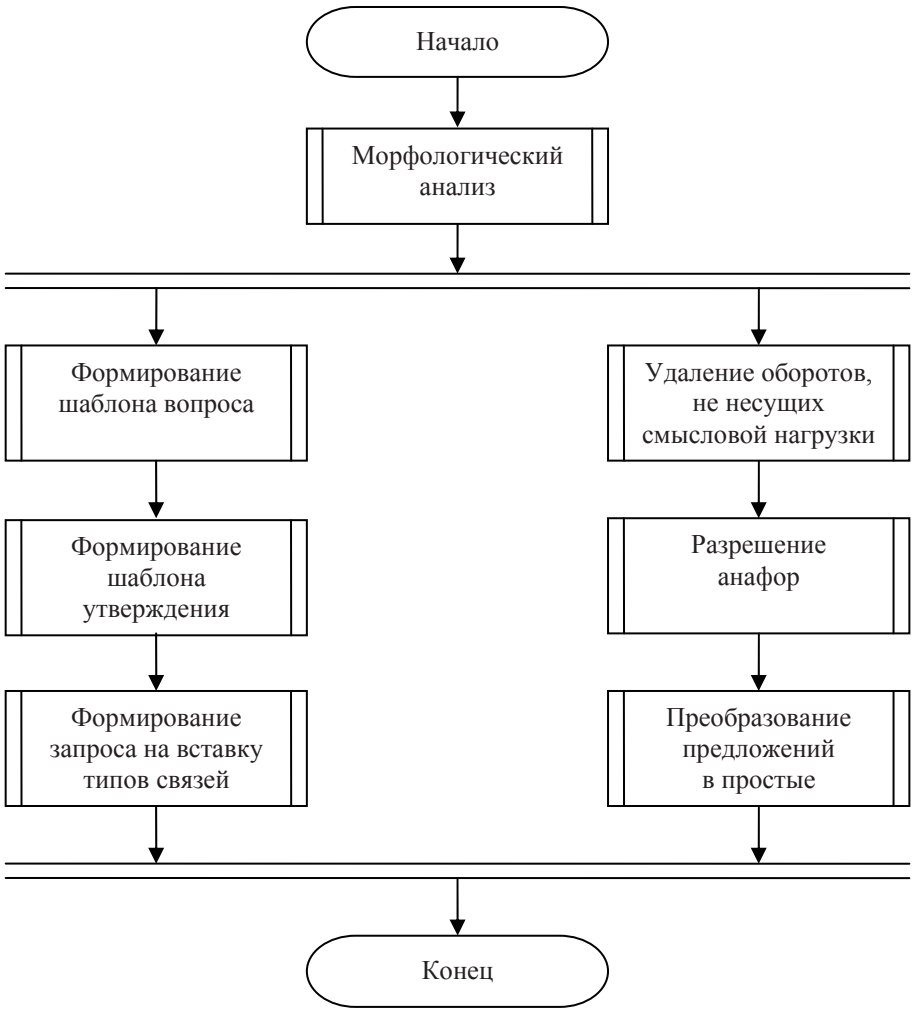

Рис. 1. Блок-схема алгоритма подготовки текста к семантическому анализу

Поскольку в текстах научного стиля встречаются предложения и с другой структурой, целесообразно создать список всех структур предложений, включив в него универсальные для всех текстов типы связей, для последующего выявления информационных единиц в ходе семантического анализа. Примером таких универсальных семантических связей являются причинно-следственные связи, которым соответствуют шаблоны предложений «*если [А], то [Б]*» и «*[А] потому, что [Б]*». Важно отметить, что части таких предложений, соответствующие тегам [А] и [Б], часто следует рассматривать отдельно как самостоятельные предложения, для которых нужно создавать шаблон при помощи описанного алгоритма. Определить случаи, когда такое распознавание необходимо, можно по наличию в данных частях предложений сказуемых, выявляемых по морфологическим признакам.

Благодаря принадлежности большинства сказуемых к определенным частям речи и морфологическим формам существует возможность автоматической генерации списка шаблонов предложений. В основе данных шаблонов лежат сказуемые с обрезанными псевдоокончаниями. Согласно нашим наблюдениям, в качестве сказуемых в текстах научного стиля наиболее часто выступают глаголы в форме третьего лица, краткие причастия и краткие прилагательные. Для обозначения псевдоокончаний глаголов используются теги [Г1], [Г2] в соответствии со спряжением. Также используется тег [Г3] для обозначения псевдоокончаний глаголов, у которых при спряжении изменяется последняя буква основы, например, «мо*жет*» – «мо*гут*». Для окончаний кратких причастий и прилагательных используется тег [О].

Разрабатываемая диалоговая информационная система, нацеленная на формулирование точных ответов на поставленные вопросы одиночными предложениями, ориентирована на выражение семантических связей определенного типа пар из шаблонов вопросительного и утвердительного предложений. Шаблон вопросительного предложения представляет собой тег вопросительного слова [В], сказуемое с тегом псевдоокончания и тег подлежащего [А]. Таким образом, обнаружение сказуемых при помощи морфологического анализа позволяет составить базу типов связей и соответствующих им шаблонов предложений. Важно отметить, что отсутствие шаблонов предложений для прочтения семантической связи в обратном направлении требует коррекции модулей семантического анализа и диалога с пользователем.

Выявленные закономерности были использованы в специальном программном средстве, выполняющем генерацию базы типов связей семантической сети по исходному тексту. Так, из предложения «*Уравнение Нернста является чрезвычайно полезным уравнением, т. к., помимо расчёта электродных потенциалов, позволяет проводить вычисления различных термодинамических характеристик веществ и реакций, протекающих в гальванических элементах*» были выявлены два сказуемых: «позволяет» и «является». После замены окончаний данных глаголов на соответствующие теги и добавления тегов, обозначающих положение других членов вопросительного и утвердительного предложений относительно сказуемого, был составлен следующий запрос на вставку в таблицу типов связей семантической сети:

INSERT INTO шаблоны VALUES

("[В] позволя[Г1] [А]", "[А] позволя[Г1] [Б]", "_", "_", "_", "", "", "_", "_", "_"),

("[В] явля[Г1]ся [А]", "[А] явля[Г1]ся [Б]", "_", "_", "_", "", "", "_", "_", "_")

В ходе ручной подготовки текстов научного стиля к семантическому анализу был выработан ряд правил преобразования предложений для их корректного актуального членения.

Первое правило заключается в удалении вводных и вставных конструкций. Данные конструкции не несут собственной смысловой нагрузки и используются только для связи семантически нагруженных фрагментов в единый текст. При формировании диалоговой системой точечных ответов из одного предложения в данных оборотах нет необходимости. Так, в предложении «*К тому же величина $E_{диф}$ существенно зависит от структуры пограничного переходного слоя, в котором распределяется диффузионный поток электролита; а структура слоя, в свою очередь, определяется способами создания и стабилизации жидкостного соединения двух растворов*» в соответствии с данным правилом удаляются обороты «*к тому же*», «*а*» и «*в свою очередь*». В результате база знаний будет пополнена семантическими связями, соответствующими частям предложения «*величина $E_{диф}$ существенно зависит от структуры пограничного переходного слоя, в котором распределяется диффузионный поток электролита*» и «*структура слоя определяется способами создания и стабилизации жидкостного соединения двух растворов*».

Кроме общих для всех текстов конструкций, к вспомогательным оборотам следует относить некоторые обороты, включающие специфические для конкретного текста глаголы. Отличить глаголы во вспомогательных конструкциях от глаголов, выражающих семантические связи, можно по морфологическим признакам: глаголы во вспомогательных оборотах,

как правило, стоят в первом лице, а глаголы, выражающие семантические связи, – в третьем. Вспомогательными оборотами могут быть целые предложения, в которых сказуемым является глагол в форме первого лица. На рис. 2 приведены результаты реализации данного правила в разработанном приложении.

Рис. 2. Результаты поиска вспомогательных конструкций в тексте

Как видно из рисунка, реализованная функция выявления вводных и вставных конструкций обладает достаточно высокой скоростью обработки текста (порядка 280 символов в секунду) и успешно выявила вводное предложение по наличию в нем глагола в форме первого лица множественного числа «*рассмотрим*». Данное предложение является примером использования распространенного в текстах научного стиля приема «авторское мы». В таких случаях в предложениях невозможно выявить подлежащее, а значит и семантическую связь, поэтому такое предложение исключается из текста при семантическом анализе.

Второе правило заключается в разрешении анафор, т. е. местоимений и других оборотов, которые обозначают информационные единицы, встречавшиеся в тексте ранее. При формировании диалоговой системой точечных ответов из одного предложения применение анафор недопустимо, т. к. перед ответным предложением отсутствует необходимый контекст. В варианте текста для семантического анализа анафорические обороты, выступающие в роли информационных единиц, должны быть заменены на полные наименования информационных единиц. Например, в соответствии с данным правилом во втором предложении пары «*Если имеется обратимо работающий ГЭ, то для него E > 0 по определению (см. подразд. 2.4). Такой элемент способен произвести максимальную по величине полезную работу по переносу электрического заряда*» следует заменить оборот «*такой элемент*» на выражение «*обратимо работающий гальванический элемент*» из первого предложения.

Третье правило заключается в преобразовании сложных предложений и предложений с однородными сказуемыми, характерных для текстов научного стиля, в простые предложения с одним сказуемым. Именно благодаря данному правилу каждая связь семантической сети будет отвечать на конкретный вопрос, не затрагивая информационные единицы, стоящие в одном сложном предложении, а также выбирая связь конкретного типа из предложений, выражающих несколько семантических связей. При форматировании текста в соответствии с данным правилом фрагменты предложений, стоящие в скобках и являющиеся частями сложносочиненных предложений, будут нуждаться в дополнении подлежащим, а иногда и сказуемым. Для уточнений в скобках нужно подбирать глагольное сказуемое из универсальных для всех текстов слов в зависимости от контекста, а не искать в тексте. Для случаев, когда в скобках находится один термин, используется глагол «*называется*», а при наличии в скобках перечисления – глагол с предлогом «*состоит из*». Так, предложение «*Знак $Q_{обр}$ совпадает со знаком ΔS реакции, протекающей в гальваническом элементе: если энтропия увеличивается, то теплота поглощается из окружающей среды*» следует разбить на два по двоеточию.

Важно отметить отсутствие необходимости в преобразовании сложноподчиненных предложений в простые. Например, в предложении «*Все химические гальванические элементы (ХГЭ) составлены из электродов, различающихся по своей химической природе*» выражение «*из электродов, различающихся по своей химической природе*» следует рассматривать как единую информационную единицу, т. к. благодаря причастному обороту выявляется новая информационная единица, отличная от термина «*электроды*».

Заключение. Описанная обработка текста позволяет приступить к автоматической генерации семантической сети [4]. Разработано программное средство, позволяющее составить базу типов связей семантической сети на основании результатов морфологического анализа. Для подготовки текста к семантическому ана-

лизу выработаны правила, позволяющие привести исходные предложения в соответствие с созданными или универсальными типами связей. Реализован автоматический поиск вспомогательных конструкций, не содержащих семантических связей, для их дальнейшего удаления.

Литература

1. Представление знаний в информационных системах. Томск: Изд-во ТПУ, 2007. 160 с.

2. Большаков И. А., Большакова Е. И. Автоматический морфоклассификатор русских именных групп // Компьютерная лингвистика и интеллектуальные технологии: материалы Междунар. конф., Бекасово, 30 мая – 3 июня 2012 г. В 2 т. Т. 1: Основная программа конференции / РГГУ. М., 2012. С. 81–92.

3. Стилистический энциклопедический словарь русского языка / под ред. М. Н. Кожиной. М.: Флинта: Наука, 2011. 696 с.

4. Гурин Н. И., Жук Я. А. Генератор семантической сети информационной системы в таблицу реляционной базы данных // Труды БГТУ. 2015. № 6: Физ.-мат. науки и информатика. С. 181–185.

References

1. *Predstavlenie znaniy v informatsionnykh sistemakh* [Knowledge representation in information systems]. Tomsk, TPU Publ., 2007. 160 p.

2. Bol'shakov I. A., Bol'shakova E. I. [Automated morphoclassifier of russian nominal groups] *Materialy Mezhdunarodnoy konferentsii (Komp'yuternaya lingvistika i intellektual'nye tekhnologii)* [Materials of the International conference (Computer linguistics and artificial intelligence)], Bekasovo, 2012, pp. 81–92 (In Russian).

3. *Stilisticheskiy entsiklopedicheskiy slovar'* [Stylistic encyclopedic dictionary]. Moscow, Flinta Publ., 2011. 696 p.

4. Gurin N. I., Zhuk Ya. A. The information system semantic network generator to a relational database table generator. *Trudy BGTU* [Proceedings of BSTU], 2015, no. 6: Physical-mathematical sciences and informatics, pp. 181–185 (In Russian).

Информация об авторах

Гурин Николай Иванович − кандидат физико-математических наук, доцент кафедры информационных систем и технологий. Белорусский государственный технологический университет (220006, г. Минск, ул. Свердлова, 13а, Республика Беларусь). E-mail: ngourine@mail.ru

Жук Ярослав Александрович − аспирант. Белорусский государственный технологический университет (220006, г. Минск, ул. Свердлова, 13а, Республика Беларусь). E-mail: zhuk@belstu.by

Information about the authors

Gurin Nikolay Ivanovich − PhD (Physics and Mathematics), Assistant Professor, the Department of Information Systems and Technologies. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: ngourine@mail.ru

Zhuk Yaroslav Aleksandrovich − PhD student. Belarusian State Technological University (13a, Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: zhuk@belstu.by

Поступила 25.04.2017