

ПОДХОД К ПРОЕКТИРОВАНИЮ СИСТЕМЫ АВТОМАТИЧЕСКОГО ПОИСКА НЕСТЫКОВОК В ГОСТАХ И СТАНДАРТАХ

Вартанян Р.А., Смольянинова В.А.

«МИРЭА – Российский технологический университет», 119454, Россия, г. Москва, проспект Вернадского, 78, e-mail: rub.1996222@gmail.com, valerysmol@mail.ru

В статье авторами предлагается подход к проектированию системы автоматического поиска нестыкровок в стандартизирующих документах. Описаны методы извлечения знаний из текста на естественном языке и построения онтологий. Приведены примеры правил для поиска противоречий в структурированном наборе требований.

Ключевые слова: стандарты, онтология, нестыковки, дескрипционная логика, извлечение знаний.

APPROACH TO DESIGNING A SYSTEM OF AUTOMATIC DISCONNECTION IN GUESTS AND STANDARDS

Vartanyan R. A., Smolyaninova V. A.

«MIREA – Russian Technological University», 119454, Moscow, 78 Vernadskogo Avenue, Russia e-mail: rub.1996222@gmail.com, valerysmol@mail.ru

In the article, the authors propose an approach to the design of a system for automatic search for inconsistencies in standardizing documents. Methods of extracting knowledge from text in natural language and constructing ontologies are described. Examples of rules for finding inconsistencies in a structured set of requirements are given.

Key words: standards, ontology, inconsistencies, descriptive logic, knowledge extraction.

Введение

В 2015 г. был принят федеральный закон N 162 «О стандартизации в Российской Федерации». В нём закрепились общие правила добровольности стандартов. Однако в статье 26.ПЗ «Общие правила применения документов национальной системы стандартизации» заявлялось, что «применение национального стандарта является обязательным для изготовителя и (или) исполнителя в случае публичного заявления о соответствии продукции национальному стандарту, в том числе в случае применения обозначения национального стандарта в маркировке, в эксплуатационной или иной документации, и (или) маркировки продукции знаком национальной системы стандартизации». Это положение кардинально изменило подход к стандартизации, поскольку сделало обязательными стандарты, заявленные заказчиком в техническом задании или производителем в рекламе.

Как следствие, возник риск ситуации, когда один стандарт ссылается на второй, второй – на третий и в результате обязательными становятся требования, противоречащие друг другу. Причина наличия стандартов, несогласованных между собой – тема отдельной статьи. Здесь ограничимся тем, что приведём примеры таких документов.

Ручная проверка всей цепочки стандартов, упомянутых в ТЗ, требует существенных затрат на этапе анализа. Очевидна необходимость автоматизации этого процесса. В этой статье предложен проект системы автоматического поиска противоречий в стандартах, законах и прочих документах.

Алгоритм работы системы следующий. Система принимает на входе пакет документов. На его основе автоматически строится OWL-онтология, описывающая предметную область. Полученная структура сканируется на наличие разного рода противоречий. На выходе система выдаёт пары несовместимых требований.

1. Примеры нестыкровок в гостах и стандартах

Перед началом проектирования авторами были найдены реальные примеры ГОСТов и стандартов, выставляющих противоречивые требования к одним и тем же объектам стандартизации. Найденные примеры были разделены на две группы: нестыковки в понятиях и нестыковки в предикатах.

1.1 Нестыковки в предикатах

Выражаются в наделении объекта свойством несочетаемым с уже определёнными характеристиками.

Например:

Из Правил устройства электроустановок (ПУЭ) издание 7 п. 6.1.21: «Аварийное освещение разделяется на освещение безопасности и эвакуационное». В то же время из свода правил обязательного применения, входящего в перечень стандартов и сводов правил к ФЗ N 384 «Технический регламент о безопасности зданий и сооружений» от 30.12.2009: «Аварийное освещение подразделяется на эвакуационное и резервное».

Из свода правил 31-110-2003 по проектированию и строительству – проектирование и монтаж электроустановок жилых и общественных зданий: «Требования о сохранении работоспособности электропроводок в условиях пожара отсутствуют». Из Статьи 82, часть 2 ФЗ N 123 «Технический регламент о требованиях пожарной безопасности» от 22.07.2008: «Кабельные линии и электропроводка систем противопожарной защиты, средств обеспечения деятельности подразделений пожарной охраны, систем обнаружения пожара, оповещения и управления эвакуацией людей при пожаре, аварийного освещения на путях эвакуации, аварийной вентиляции и противодымной защиты, автоматического пожаротушения, внутреннего противопожарного водопровода, лифтов для транспортировки подразделений пожарной охраны в зданиях и сооружениях должны сохранять работоспособность в условиях пожара в течение времени, необходимого для выполнения их функций и эвакуации людей в безопасную зону».

1.2 Нестыковки в понятиях

Выражаются в переопределении понятия, которому уже дано определение.

Например:

Из свода правил обязательного применения 52.13330.2011 Естественное и искусственное освещение помещений вновь строящихся и реконструируемых зданий и сооружений различного назначения, входящего в перечень стандартов и сводов правил к ФЗ N 384 «Технический регламент о безопасности зданий и сооружений» от 30.12.2009, приложение Б: «эвакуационный выход: выход, предназначенный для эвакуации людей при чрезвычайных ситуациях». Из ГОСТ Р 55842-2013 «Освещение аварийное. Классификация и нормы», пункт 3.8: «эвакуационный выход: выход, предназначенный для использования в аварийной ситуации». Одновременно с этим из Положения 48, Статьи 2 ФЗ N 123 «Технический регламент о требованиях пожарной безопасности» от 22.07.2008: «эвакуационный выход – выход, ведущий на путь эвакуации, непосредственно наружу или в безопасную зону».

Из свода правил СП 118.13330.2012 «Общественные здания и сооружения» (действует с 01.09.2014г.) Приложении Б пункт 20: «помещение с массовым пребыванием людей: Помещение с количеством людей более 1 чел. на 1 м² площадью 50 м² и более». В то время как в Постановлении Правительства от 25.04.2012 г. N 390 «О ПРОТИВОПОЖАРНОМ РЕЖИМЕ» (в ред. Постановления Правительства РФ от 17.02.2014 N 113): «на объектах, кроме жилых домов, на которых может одновременно находиться 50 и более человек, то есть с массовым пребыванием людей, руководитель организации может создавать пожарно-техническую комиссию». То есть Постановление Правительства расширяет список помещений, относимых к помещениям с массовым пребыванием людей, поскольку не учитывает площадь объекта.

Отдельно выделяют как нестыковку использование термина, которому ещё не дано определение. Так, в уже упомянутой Статье 82, часть 2 ФЗ «Технический регламент о требованиях пожарной безопасности» при описании требований к элементам системы противопожарной защиты устанавливают временные ограничения равные «времени, необходимому для выполнения их функций и эвакуации людей в безопасную зону». При этом ни в самом законе, ни в других постановлениях не конкретизируется ни примерные рамки, ни способ вычисления этого параметра.

2. Поиск нестыковок в онтологии

Авторами предлагается осуществлять поиск нестыковок не напрямую в исходном тексте документа, написанном на естественном языке, а в онтологии, созданной на его основе и пригодной для машинной обработке. Схема представления данных в виде онтологии удобна тем, что позволяет наделить объект определённого типа фиксированным набором атрибутов. Исходя из наполненности атрибутов значениями, можно судить о полноте предоставляемых об объекте сведений. Отношения между объектами также рассматриваются как атрибуты, значение которых – другие объекты.

В соответствии классификацией противоречий, приведённой выше в пункте 1, авторами были сформулированы правила поиска нестыковок разного типа в онтологиях. За основу была взята построенная на основе текста онтология, обозначенная как От.

2.1 Поиск нестыковок в предикатах

В формальном представлении предикат отражает суждение, высказанное о каком-либо объекте, в виде отношения. Исходя из этого, нестыковкой в предикатах можно считать присутствие взаимоисключающих отношений. Виды противоречий в предикатах – в таблице 1.

Таблица. 1. Виды нестыкровок в предикатах

Вид	Описание
Расширение	Объект X связан через отношение L с объектами Y и Z. При этом отношение L может встречаться у одного объекта только один раз.
Прямое отрицание	Объекты X и Y вступают в отношения L и вместе с тем отношение L декларируется как невозможное между X и Y.

В качестве примера расширения как вида нестыковки подойдёт пара утверждений из п. 1.1: (1) «Аварийное освещение разделяется на освещение безопасности и эвакуационное», (2) «Аварийное освещение подразделяется на эвакуационное и резервное». Чтобы показать противоречивость утверждений, приведём семантический анализ этих предложений.

Результатом семантического анализ является структура в форме дерева, где слова выступают в роли семантических узлов, связанных друг с другом посредством семантических отношений. Проектом Автоматическая Обработка Текста [4] разработан Dialing Graph Builder для семантического анализа отдельных предложений длиной до 150 символов. Результат обработки предложений 1-2 – на рисунке 1.

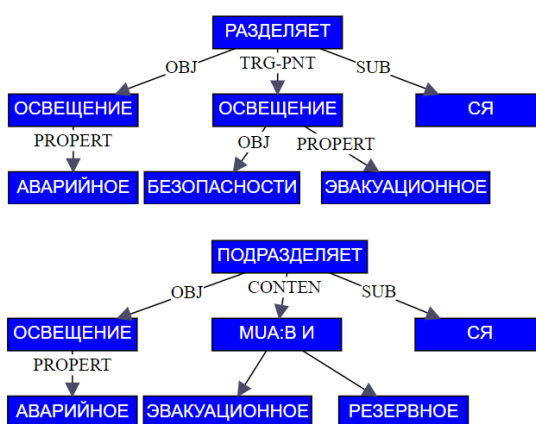


Рис. 1. Семантический разбор предложений 1-2

Как видно из рисунка 1, объект (OBJ) «ОСВЕЩЕНИЕ» вступает в связку с «РАЗДЕЛЯЕТ» и «ПОДРАЗДЕЛЯЕТ». Условимся, что «РАЗДЕЛЯЕТ(СЯ)» и «ПОДРАЗДЕЛЯЕТ(СЯ)» – контекстные синонимы слову «КЛАССИФИЦИРУЮТСЯ». Тогда, видим, что в онтологии приведены две классификации. В обеих фигурирует тип «ЭВАКУАЦИОННОЕ». Это исключает вариант, что классификации основаны на разных признаках.

Пример прямого отрицания: (3) «Требования о сохранении работоспособности электропроводок в условиях пожара отсутствуют», (4) «Кабельные линии и электропроводка... должны сохранять работоспособность в условиях пожара». Разбор – на рисунке 2.

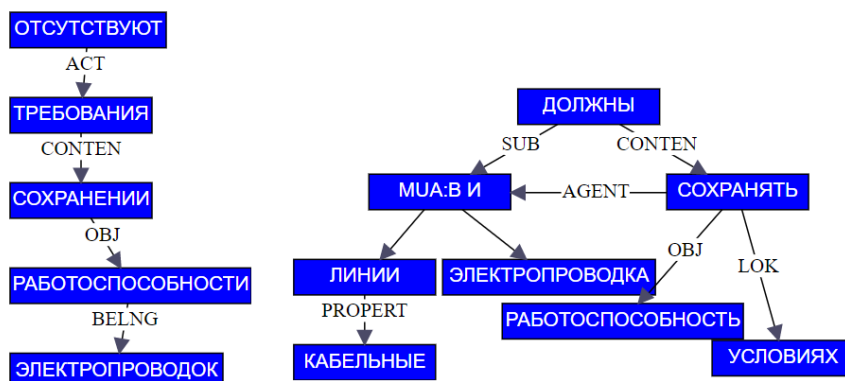


Рис. 2. Семантический разбор предложений 3-4

Здесь видно, что семантические вершины предложений – «ОТСУТСТВУЮТ» [«ТРЕБОВАНИЯ»] и «ДОЛЖНЫ» – это контекстные антонимы.

2.2 Поиск нестыкровок в понятиях

Пусть в онтологии *Ot* существуют понятия *X* и *Y*. Возможные виды нестыкровок представлены в таблице 2.

Таблица 2. Виды нестыкровок в понятиях

Вид	Описание
Дублирование. Понятие <i>X</i> дублирует своим определением понятие <i>Y</i> .	В онтологии <i>Ot</i> одновременно существуют понятия <i>X</i> и <i>Y</i> . При этом понятие <i>X</i> выражено словом или словосочетанием <i>Z</i> , которое также используется для определения понятия <i>Y</i> .
Использование термина, которому ещё не дано определение	В онтологии <i>Ot</i> существует понятия <i>X</i> , определённое через понятие <i>Y</i> , которому ранее не дано определения.

В качестве примера дублирования понятия подойдёт пара утверждений из пункта 1.2: (5) «эвакуационный выход: выход, предназначенный для эвакуации людей при чрезвычайных ситуациях», (6) «эвакуационный выход: выход, предназначенный для использования в аварийной ситуации». Их разбор – на рисунке 3.

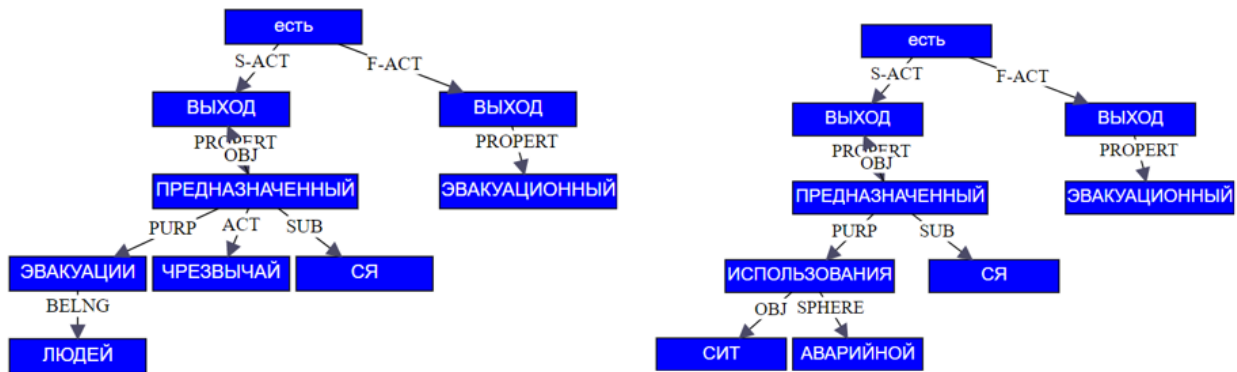


Рис. 3. Семантический разбор предложений 5-6

В исходных предложениях нет слова «есть», но при построении семантической структуры «ЕСТЬ» добавляется, чтобы связать термин и определение в виде цепочки: «ТЕРМИН» – F-АКТ – «ЕСТЬ» – S-АКТ – «ОПРЕДЕЛЕНИЕ».

Пример использования термина, которому ранее не дано определение: (7) «Кабельные линии и электропроводка... должны сохранять работоспособность... в течение времени, необходимого для выполнения их функций и эвакуации людей в безопасную зону». На дереве структуры предложения 7 (рис. 4) видно, что у вершины «ВРЕМЯ» нет связи F-АКТ с «ЕСТЬ» с последующим ответвлением S-АКТ от «ЕСТЬ», характерной для предложений, раскрывающих значение термина.

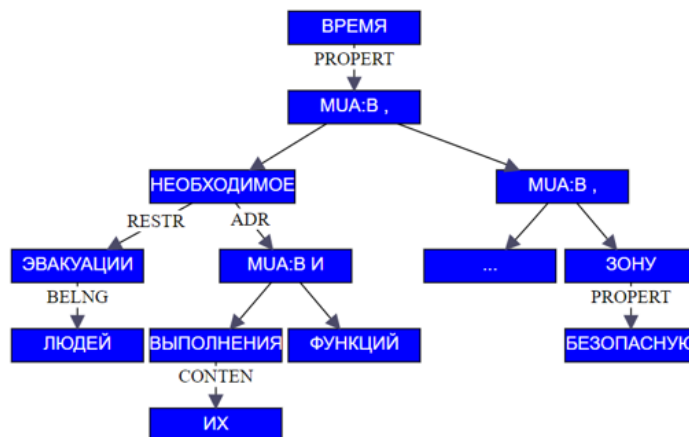


Рис. 4. Семантический разбор предложения 7

3. Построение онтологии

Цель построения онтологии – представить содержание текста в формальном виде, пригодном для дальнейшей машинной обработки. Под «формальным видом» понимается модель, состоящая из классов, объектов и свойств [1].

Для описания онтологии нормативно-правовых документов будем использовать язык дескрипционной логики OWL 2 DL (Web Ontology Language 2 Description Logic). Его преимущества: выразительность, достаточная для формулировки практически значимых фактов, разрешимость (вычисления завершаются за конкретное время), наличие алгоритмов логического анализа для вывода новых знаний из уже имеющихся и присутствие в открытом доступе уже реализованных систем, осуществляющих логический вывод.

Автоматическое построение онтологии на основе естественного текста представляет собой задачу из области выделения знаний на основе онтологий (англ. Ontology-based information extraction, OBIE)[2]. Её сложность заключается в необходимости разработки инструментов распознавания понятий, сущностей и отношений на языке самого текста. Большинство существующих средств работает с текстами на английском языке. Конечно, программы-переводчики могут быть использованы для перевода исходных текстов на язык, пригодных для построения онтологии, но в связи со сложностью делового стиля изложения стандартизирующих документов эффективность переводчиков низка.

Среди русскоязычных OBIE-инструментов стоит выделить ABBYY Compeno и продукты группы aot.ru. Их принцип работы схож: поступающие на вход предложения подвергаются семантическому анализу с последующей идентификацией типа связи между сущностями. Оба инструмента способны распознавать десятки видов связей: «Автор-действие», «Событие-дата», «Объект-место» и т.д.[3]

ABBYY Compeno – исключительно коммерческий продукт. Он был разработан для внутреннего пользования при создании решений интеллектуального поиска и виртуальных собеседников. Инструменты AOT есть в свободном доступе в демо-режиме, ограничивающим размер обрабатываемых предложений. Полный доступ возможен через технологию REST API по договорённости с создателями.

4. Архитектура системы поиска нестыкочков

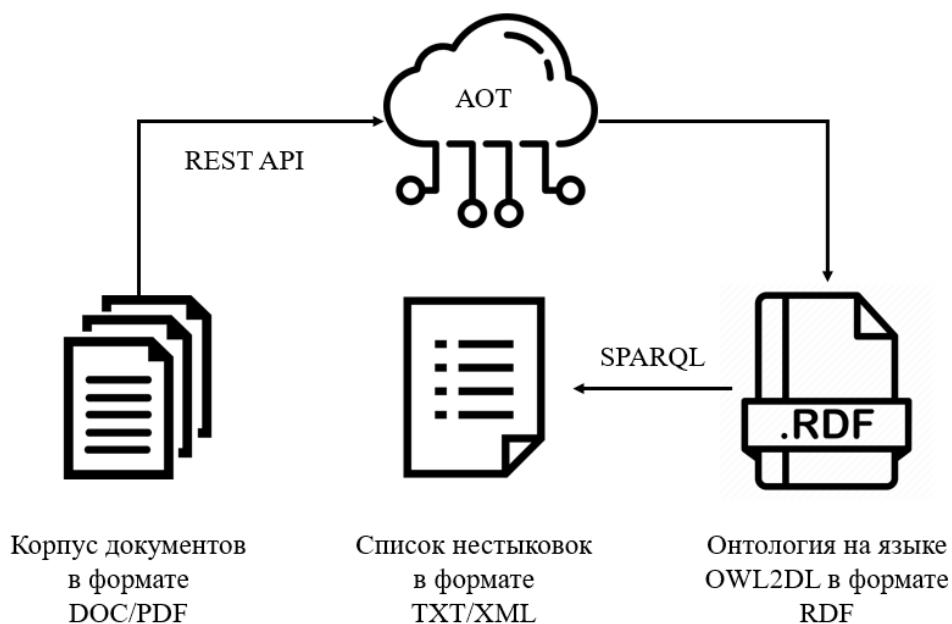


Рис. 5. Архитектура системы поиска нестыкочков

На рисунке 5 изображена предложенная архитектура системы автоматического поиска нестыкочков в текстах стандартизирующих документов. На входе система получает пакет документов в формате DOC либо PDF. Файлы отправляются на сервер Автоматической Обработки Текста (АОТ) посредством REST API. В ответ на запрос АОТ формирует онтологию на основе совокупности полученных текстов на языке OWL2DL.

Для поиска нестыкровок к полученной онтологии делается запрос на языке SPARQL. Выбор в пользу SPARQL в качестве языка запросов обоснован рекомендацией консорциума W3C [5]. В одном запросе отражены все правила поиска нестыкровок, изложенные в пунктах 3.1-3.2. Также предполагается возможность отдельного извлечения нестыкровок разного вида, например, только в понятиях и только типа «дублирование».

На выходе система выдаёт массив пар, противоречащих друг другу утверждений.

Заключение

В работе предложена схема классификации нестыкровок и приведены формальные правила выявления разного рода противоречий. Описан подход к проектированию системы автоматического поиска нестыкровок, в основе которого лежит отечественная разработка по построению онтологий.

Список литературы

1. Добров Б. В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. Онтологии и тезаурусы: модели, инструменты, приложения. — М.: Бинوم. Лаборатория знаний, 2009. — 173 с.
2. Daya C. Wimalasuriya, Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches // Journal of Information Science. — 2010. — Т. 36, вып. 3. — С. 306 – 323.
3. Илья Булгаков, Алгоритм извлечения информации в АBBYY Compreno. Часть 1. Режим доступа: <https://habr.com/ru/company/abbyy/blog/269191/>
4. Леонтьева Н., Панкратов Д., Сокирко А., Кобрицов Б. Автоматическая обработка текста [Электронный ресурс]. – [2008]. – Режим доступа: <http://www.aot.ru/technology.html>
5. Semantic Web Activity News (англ.). — Новости семантической паутины. Режим доступа: https://www.w3.org/blog/SW/2008/01/15/sparql_is_a_recommendation/

References

1. Dobrov B.V., Ivanov V.V., Lukashevich N.V., Soloviev V.D. Ontologies and thesauri: models, tools, applications. - M.: Binom. Knowledge Laboratory, 2009. -- 173 p.
2. Daya C. Wimalasuriya, Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches // Journal of Information Science. - 2010. - T. 36, no. 3. - S. 306 - 323.
3. Ilya Bulgakov, Information Extraction Algorithm in ABBYY Compreno. Part 1. Access mode: <https://habr.com/ru/company/abbyy/blog/269191/>
4. Leontyeva N., Pankratov D., Sokirko A., Kobritsov B. Automatic text processing [Electronic resource]. - [2008]. - Access mode: <http://www.aot.ru/technology.html>
5. Semantic Web Activity News (English). - Semantic Web News. Access mode: https://www.w3.org/blog/SW/2008/01/15/sparql_is_a_recommendation/