УДК 004.021

ИССЛЕДОВАНИЕ МЕТОДОВ ИЗВЛЕЧЕНИЯ СЕМАНТИЧЕСКИХ РОЛЕЙ ИЗ ТЕХНИЧЕСКИХ ТРЕБОВАНИЙ

Турыгин Д.О. (Университет ИТМО)
Научный руководитель – доцент, кандидат технических наук, Баймуратов И.Р.
(Университет ИТМО)

Введение. В докладе представлено исследование методов извлечения семантических ролей из требований нормативно-технической документации, а также оценивается их применимость к текстам на русском языке. На сегодняшний день такие требования представлены, как правило, в слабоструктурированном человекочитаемом формате, из-за чего проверка информационных моделей на соответствие требованиям возможна только в ручном режиме. Первым шагом при формализации технических требований может быть разметка семантических ролей, но такая разметка потребует от пользователя большого количества времени и квалификации в вопросах семантики. Таким образом, данное исследование направлено на автоматизацию разметки (извлечение) семантических ролей в технических требованиях.

Основная часть. В ранее опубликованной нами статье был предложен формат разметки технических требований [1]. Согласно этому формату, семантические роли соответствуют структуре утверждений дескрипционной логики и включают в себя субъект и объект. Роль “субъект” соответствует левой части утверждения. Это класс объектов, к которому относятся ограничения, описанные в требовании, например, “лестница”, “дверь” и т.д. В свою очередь, роль “объект” соответствует правой части утверждения и содержит ограничения, выраженные в требовании и накладываемые на субъект, например, “должна быть ровной, без выступов”.

При анализе методов извлечения семантических ролей из требований нормативно-технической документации рассматривается 2 подхода. Первый основан на синтаксических деревьях, второй – на деревьях зависимостей [2]. Синтаксические деревья – это деревья, формируемые на основе POS-тегов (part-of-speech). В дереве зависимостей каждый узел представляет слово в предложении, а каждая дуга отражает грамматическую зависимость между двумя словами. Первый метод проблематично применить к нашей задаче ввиду того, что на сегодняшний день нет ни одного инструмента, который позволял бы строить синтаксические деревья для русского языка. Ввиду специфики построения предложений в русском языке, нет возможности создания универсальных правил построения такого дерева [3].

Подход на основе дерева зависимостей является более перспективным. Узлы индексируются в соответствии с порядком слов в предложении. Корень дерева разделяет предложение на две части: левую, соответствующую субъекту, и правую, соответствующую объекту. К левой части относятся ветви, индекс первого узла которых меньше индекса корня. Если таких веток нет, то рассматривается первая ветвь с минимальным индексом после корня. Остальные ветви, включая сам корень, относятся ко второй части. Рассмотрим пример: “*При перепаде высот в здании или сооружении следует предусматривать лестницы, пандусы или иные подъемные*

устройства". В этом случае корневое слово – "*следует",* субъект (левая ветвь) – "*При перепаде высот в здании или сооружении"*, а объект (правая ветвь и корень) – *"следует предусматривать лестницы, пандусы или иные подъемные устройства"*.

Для реализации метода используется библиотека spaCy [4]. Для оценки точности метода был составлен набор технических требований из области строительства, они были размечены при участии экспертов в предметной области. После разметки требования разбивались на предложения. Из них были отобраны только те предложения, которые содержат ровно один субъект и один объект. Таким образом была получена выборка из 44 предложений. Точность полного совпадения получившейся разметки относительно образцовой составляет 64.44%, в свою очередь, нормализованная взаимная информация равна 80.79%.

Выводы. Проведен анализ методов извлечения семантических ролей из требований нормативно-технической документации. Реализован метод на основе дерева зависимостей. Оценена точность результатов извлечения.

Список использованных источников:

1. Исследование методов разметки семантических ролей для информационного моделирования нормативных требований на языке OWL // Сборник трудов XI Конгресса молодых ученых (Санкт-Петербург, 4-8 апреля 2022 г.) – 2022. – Т. 1. – С. 451-453.
2. Anantharangachar, R., Ramani, S., & Rajagopalan, S. . Ontology Guided Information Extraction from Unstructured Text. International journal of Web & Semantic Technology – 2013.
3. Saber, Y. M., Abdel-Galil, H., & El-Fatah Belal, M. A. Arabic ontology extraction model from unstructured text. Journal of King Saud University - Computer and Information Sciences, 34 (8, Part B), – 2022. – 6066–6076.
4. spaCy // Industrial-strength Natural Language Processing in Python. // https://spacy.io