УДК 004.912

ИСПОЛЬЗОВАНИЕ КОНТЕКСТНЫХ ОПЕРАТОРОВ В ПАРАМЕТРИЗОВАННЫХ ШАБЛОНАХ ДЛЯ ОБРАБОТКИ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Савёнок В.А.
УО «Белорусский государственный университет информатики и радиоэлектроники», г. Минск, Республика Беларусь, savionak@gmail.com

В первом приближении к основным конструкциям параметризованных шаблонов поиска в тексте на естественном языке можно отнести следующие элементы:

– текстовый литерал, сравниваемый с учётом или без учёта регистра символов;

– последовательность со строгим следованием элементов;

– альтернатива (вариация), представляющая собой перечисление элементов, допустимых в данной точке шаблона;

– повторение элемента шаблона определенное число раз в рамках допустимого диапазона минимального и максимального количества;

– следование элементов через слова с указанием максимального и минимального допустимого количества слов между элементами.

Однако перечисленные конструкции не позволяют учитывать при поиске существенную составляющую текста на естественном языке – контекст [1].

Учёт контекста при проведении поиска в тексте на естественном языке позволяет расширить пространство поиска за границы синтаксиса, что повышает выразительную мощность параметризованных текстовых шаблонов и, следовательно, их точность.

Принимая во внимание высокую трудоемкость задачи семантического анализа текста, для учёта контекста выделим простейшие контекстные отношения между парой выражений A и B:

– отношение владения (вложенности) – A «*содержит*» B;

– отношение подчиненности – A «*находится внутри*» B;

– отношение исключения – A «*не содержит*» B.

Последнее из перечисленных отношений имеет более общий аналог, который определяет исключение пересечений между совпадениями выражений: A «*не пересекает*» B. В таблице 1 приведены операторы, которые ставятся в соответствие каждому из отношений.

Таблица 1 – Основные контекстные отношения и соответствующие операторы

№	**Контекстное отношение**	**Семантика**	**Оператор**
1	Совпадение выражения A содержит совпадение выражения B	«*содержит*»	A @having B
2	Совпадение выражения A содержится внутри совпадения выражения B	«*находится внутри*»	A @inside B
3	Совпадение выражения A не содержит ни одного совпадения выражения B	«*не содержит*»	– (частный случай оператора 4)
4	Совпадение выражения A не пересекается с совпадениями выражения B	«*не пересекает*»	A @outside B

В качестве примера применения контекстных операторов рассмотрим процесс повышения точности классификатора текстов на естественном языке, основанного на параметризованных шаблонах [2].

Пусть необходимо построить классификатор текстов, содержащих описания вакансий. Обозначим заданное множество классов, состоящее из двух элементов: «Transportation» и «Others». К первому классу относятся тексты, содержащие описания вакансий водителя (легкового автомобиля, грузового автомобиля, автобуса, такси и т. д.), ко второму классу – все

остальные тексты (см. рис. 1 а). Результатом классификации должны стать два множества T_1 и T_2, содержащие тексты, которые относятся к классам «Transportation» и «Others» соответственно [2].

Общий вид результата классификации представлен на рисунке 2 б.

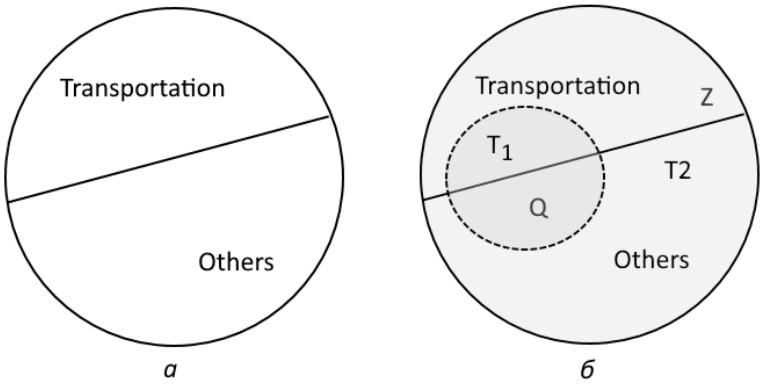

Рисунок 1 – Классы для разделения текстов (а) и предварительный результат классификации (б)

Здесь Q – элементы класса «Others», которые по результату классификации по ошибке были отнесены к классу «Transportation», Z – элементы множества «Transportation», которые по ошибке были отнесены к классу «Others». Повышение точности классификатора представляет собой уменьшение мощности множеств Q и Z. Так для исключения элементов множества Q из множества T_1 необходимо уточнить контекст употребления ключевых выражений уже составленных шаблонов. Это достигается путём применения контекстных операторов. Например, из совпадений можно исключить упоминание профессии водителя в контексте управляющих должностей (менеджер, глава отдела и т. д.) или должности инструктора:

<исходный шаблон> @outside (Предложение @having {"инструктор", "менеджер"}).

Данный псевдо-шаблон отражает исключение из рассмотрения предложений, содержащих слова «инструктор» или «менеджер», при помощи контекстных операторов «*содержит*» (@having) и «*не пересекает*» (@outside).

Список использованных источников

1. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний / Загоруйко, Н. Г // Издательство: Новосибирск: ИМ СО РАН, 1999. – 270 с.

2. Савёнок, В.А. Построение классификатора текстов на естественном языке с использованием параметризованных шаблонов / В.А. Савёнок, С.А. Медведев, В.Н. Селедец // Материалы международной научной конференции «Информационные технологии и системы 2019»: сб. статей. – Минск, 2019. – С. 268-269.