

Формализация математических текстов через теорию дискурса

С.Л. Хасанянова, А.Ю. Перепечко

Московский физико-технический институт
(национальный исследовательский университет)

Первая часть данной работы заключалась в том, чтобы разобраться, что из себя представляют синтаксические деревья разбора для естественного языка. Была изучена соответствующая теория. Затем были изучены существующие деревья разбора для формул. Например, была рассмотрена статья Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale [1]. Авторы статьи решают следующую проблему: при использовании математических формул для поиска (запрос по выражению), пригодность найденных формул часто зависит больше от идентичности символов и расположения, чем от глубокой математической семантики. Используя дерево расположения символов для представления формул, авторы статьи предлагают метод максимального сходства поддеревьев для ранжирования формул на основе подвыражений, символы и расположение которых лучше всего соответствуют формуле запроса.

Следующий шаг работы заключался в том, чтобы разобраться в понятии теории дискурса. Было рассмотрено, из каких элементов состоит описание теории дискурса. По своей сути, теория дискурса – это логика, используемая для описания (семантического содержания) естественного языка. Ее элементарной единицей является структура представления дискурса [2]. Например, такая:

x	y
$John(x)$	
$dog(y)$	
$owns(x, y)$	

Эта структура представления дискурса представляет семантическое содержание предложения

John owns a dog.

и соответствует следующей формуле в предикатном исчислении первого порядка:

$$\exists x. \exists y. (John(x) \wedge dog(y) \wedge owns(x, y))$$

Следующие шаги работы заключались в том, что были приведены примеры текста и формул, выписанных в теории дискурса. Была предпринята попытка взять дерево разбора для естественного языка и дерево разбора для формул и соединить их.

Литература

1. Richard Zanibbi, Kenny Davila, Andrew Kane, Frank Wm. Tompa. Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale.
2. Mohan Ganesalingam. The Language of Mathematics.