УДК 004.9

НОРМАЛИЗАЦИЯ ЧИСЕЛ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Усков И.В. (Национальный исследовательский университет ИТМО), **Письмак А.Е.**
(Национальный исследовательский университет ИТМО)
Научный руководитель – доцент, кандидат технических наук Перминов И.В.
(Национальный исследовательский университет ИТМО)

В данной работе рассмотрено два подхода к решению задачи нормализации числа на естественном языке — переводу числа, выраженного текстом, в его значение. Первый подход представляет собой алгоритм, основанный на грамматике русского языка, второй основан на использовании алгоритмов машинного обучения.

Введение. В обработке текста на естественном языке стоит задача нормализации числовых данных: преобразования чисел из текстовой формы в их значение. Существуют решения для обратной задачи: преобразования числа в текстовую форму, единого же подхода к нормализации чисел из текстовой формы не существует, хотя наработки в этой области есть: так, поисковые машины активно используют алгоритмы машинного обучения, которые позволяют частично решить эту задачу. Также существует программное обеспечение с открытым исходным кодом, в котором заявлена такая возможность. Но всё оно является неприменимым: в нём либо полностью отсутствует поддержка работы с русским языком, либо она является частичной.

Основная часть. В данной работе предложено два подхода к решению задачи нормализации применительно к русскому языку. Первый подход основан на грамматических правилах: числа в русском языке чаще всего выражаются числительными, которые имеют строго определённую структуру. Алгоритм получает на вход список текстовых токенов, являющиеся составной частью числа. Далее подходящие токены группируются и преобразуются в числовое значение в соответствии с правилами. Например, составное порядковое числительное состоит из набора токенов, соответствующих количественным числительным в именительном падеже единственного числа и одного токена количественного числительного, который может изменяться по своим грамматическим признакам (число, род, падеж). Дроби же состоят из числителя, выраженного количественным числительным, и знаменателя, являющегося порядковым числительным. Но такой алгоритм не всегда может однозначно произвести нормализацию. Примером является число "сто двадцать третьих", которое не сможет однозначно интерпретировать ни алгоритм, ни человек. В связи с вероятностью такой неоднозначности алгоритм на выходе даёт все возможные интерпретации числа ($\frac{120}{3}$ и $\frac{100}{23}$ в предыдущем примере). Однозначное разрешение подобных ситуаций является вопросом дальнейшего исследования.
Второй подход основан на использовании алгоритмов машинного обучения. Для реализации предыдущего алгоритма был собран набор числительных и их грамматических признаков, что позволяет сгенерировать основу для тренировочного набора данных: случайно сгенерированное целое или дробное число можно превратить в текст, следуя набору правил русского языка. Такой подход даёт единственный вариант нормализации числа.

Выводы. Для проверки работоспособности предложенных методов нормализации было сгенерировано тестовое покрытие на 10000 вариантов тестовых представлений чисел. Варианты нормализации каждого тестового образца, полученные первым алгоритмом, гарантированно содержат правильное решение, но на данный момент отсутствует возможность определить наиболее вероятное из них. Кроме того, алгоритм не покрывает некоторые широко используемые в речи конструкции, такие как подразумеваемые, но не

используемые явно слова (пропущенное слово "тысячи" в выражении "две семьсот"). Подход, основанный на использовании машинного обучения, избавлен от этого недостатка, но т.к. он даёт единственный вариант нормализации числа, он не во всех случаях оказывается верным.

Письмак А.Е. (автор)
Усков И.В. (автор)

Перминов И.В. (научный руководитель)