ПОСТРОЕНИЕ КЛАССИФИКАТОРА ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ ПАРАМЕТРИЗОВАННЫХ ШАБЛОНОВ

Савёнок В. А., Медведев С. А., Селедец В. Н.
Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: savionak@gmail.com, vadik446s@gmail.com, msa@bsuir.by

В статье приводится постановка задачи классификации текстов на естественном языке с применением параметризованных шаблонов; описывается построение классификатора с использованием параметризованных шаблонов на примере классификации текстов, содержащих описание вакансий. Также приводится пример оценки точности построенного классификатора и подходы к ее повышению.

Введение

Задача классификации текстов на сегодняшний день широко распространена. В общем случае, для того, чтобы точно классифицировать множество текстов, необходимо проанализировать их содержимое на предмет наличия ключевых слов и фраз, которые определяют принадлежность данных текстов к заданным классам. Подход с использованием параметризованных шаблонов позволяет эффективно осуществить данный анализ и построить классификатор, обладающий достаточно высокой точностью.

I. Постановка задачи

При применении параметризованных шаблонов для решения задачи классификации множества текстов T на множество классов C, необходимо составить такое множество шаблонов P, чтобы множество возможных совпадений M_i, порождаемых шаблоном P_i, однозначно принадлежало множеству текстов, образующих класс C_i (см рис. 1).

Другими словами, каждому классу ставится в соответствие целевой шаблон. Наличие совпадения целевого шаблона в тексте говорит о принадлежности данного текста к классу, которому был поставлен в соответствие данный целевой шаблон. Такая формулировка решения позволяет определять принадлежность текста к нескольким классам [1].

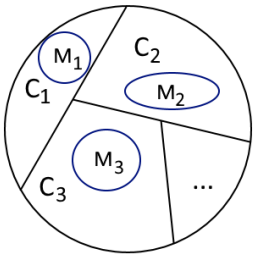

Рис. 1 – Принадлежность множеств совпадений к классам

II. Построение классификатора текстов на основе параметризованных шаблонов

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 (грузового автомобиля, легкового автомобиля, такси, автобуса и т. д.), а ко второму классу – все остальные тексты (см. рис. 2 а). Результатом классификации должны стать два множества T_1 и T_2, содержащие тексты, которые относятся к классам «Transportation» и «Others» соответственно. Общий вид результата классификации представлен на рисунке 2 б.

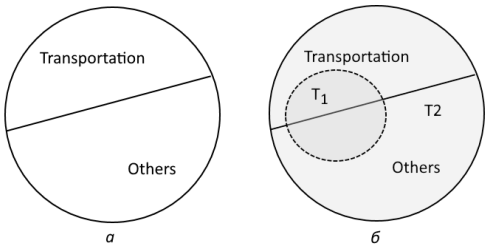

Рис. 2 – Классы для разделения текстов (а) и предварительный результат классификации (б)

Для решения поставленной задачи достаточно составить один параметризованный шаблон, такой, чтобы все совпадения, порождаемые им, принадлежали множеству текстов, относящихся к классу «Transportation». Тексты, в которых не будут найдены совпадения указанного шаблона, будут отнесены к классу «Others».

Множество возможных совпадений, которое должно порождаться целевым шаблоном, определяется путем анализа предметной области. Элементами данного множества выступают ключевые, специфические для данной предметной области, слова и фразы, которые употребляются в определенном контексте. Для текстов, содержащих описания вакансий, ключевыми словами являются названия профессий. Следовательно, они являются частью множества совпа-

дений, которые должен порождать целевой шаблон. Такой целевой шаблон позволит определить множество текстов T_1, принадлежащих классу «Transportation». Все оставшиеся тексты составят множество текстов T_2, относящихся к классу «Others».

III. Оценка точности классификатора

Для определения готовности классификатора к решению поставленной задачи необходимо оценить его точность. Способ определения точности классификатора зависит от типа входных данных, которые могут являться как обучающей выборкой, так и уже классифицированными реальными данными.

Принимая во внимание тот факт, что для обучающей выборки количество классов и количество текстов в каждом классе известны заранее, точность классификатора можно рассчитать по формуле

$$a = \frac{\sum_{i=1}^{N} c_i}{\sum_{i=1}^{N} n_i},$$

где N – количество заданных классов, c_i – число текстов, которые отнесены классификатором к классу i; n_i – общее число текстов, принадлежащих классу i [2].

При создании классификатора на основании реальных данных, в которых тексты уже отнесены к одному или нескольким классам некоторым классификатором L_1, заранее известны и количество классов, и количество текстов, принадлежащих данным классам. Однако, если классификатор L_1 не предоставляет требуемой точности, возникает необходимость в создании нового классификатора L_2 на основании полученных данных.

В силу того, что классификатор L_1 не является абсолютно точным, не все тексты, отнесенные к заданным классам, на самом деле принадлежат им. Для расчета точности создаваемого классификатора L_2 можно воспользоваться следующей формулой:

$$a = \frac{|Q| + |U_{L_2}|}{|F| + |U_{L_1}|},$$

где Q – множество текстов, для которых класс определен верно обоими классификаторами; U_{L_k} – множество текстов, класс которых определен верно классификатором L_k, при этом $U_{L_k} \cap Q = \emptyset$, $k = 1, 2$; F – множество текстов, обработанное классификатором L_2 [2].

IV. Повышение точности классификатора

Как видно из рисунка 2 б, полученное после предварительной классификации множество T_1 не является полным, так как не все тексты класса «Transportation» вошли в T_1. Кроме того, T_1 содержит тексты, принадлежащие классу «Others», что требует дальнейшего уточнения целевого шаблона. Для повышения точности классификатора необходимо решить две задачи (см. рис. 3):

1. исключить из T_1 тексты, принадлежащие классу «Others»;
2. включить в T_1 оставшиеся тексты, принадлежащие классу «Transportation».

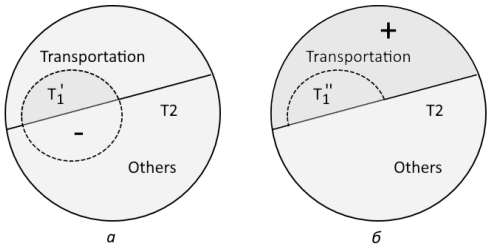

Рис. 3 – Уточнение результатов классификации путем исключения (а) и включения (б) текстов, принадлежащих соответствующим классам

Для исключения из T_1 текстов, принадлежащих классу «Others», необходимо проанализировать эти тексты и добавить в целевой шаблон конструкцию исключения, которая содержит ключевые слова и фразы, не относящиеся к «Transportation». Также необходимо уточнить контекст употребления уже заданных ключевых выражений. Для этого можно использовать параметризованный шаблон, совпадения которого будут состоять из заданного ключевого слова или фразы с захватом окружающих слов. Полученный шаблон необходимо применить к текстам, которые ошибочно попали во множество T_1, после чего найденные совпадения следует добавить в качестве исключений в целевой шаблон.

Для расширения множества T_1 необходимо проанализировать тексты, ошибочно не отнесенные к классу «Transportation», и дополнить ключевые слова и фразы новыми элементами.

Значительное повышение качества классификатора достигается путем выполнения описанных операций в несколько итераций.

Заключение

Описанная методика построения классификатора текстов с использованием параметризованных шаблонов была успешно применена на практике для выявления текстов на английском языке, содержащих описание вакансии водителя. Достигнутая точность классификатора на выборке текстов из открытых источников [3] по полученным оценкам составляет более 91%.

1. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний / Загоруйко, Н. Г // Издательство: Новосибирск: ИМ СО РАН, 1999. – 270 с.
2. Оцека классификатора (точность, полнота, F-мера) [Электронный ресурс] – Режим доступа: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html. – Дата доступа: 08.09.2019.
3. Driver Jobs, Employment | Indeed.com [Электронный ресурс] – Режим доступа: https://www.indeed.com/q-driver-jobs.html. – Дата доступа: 15.09.2019.